UNIMORE
UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA

Dipartimento di Economia Marco Biagi

DEMB Working Paper Series

N. 170

Nonresponse and measurement errors in income: matching individual survey data with administrative tax data

Michele Lalla[1], Maddalena Cavicchioli[2]

March 2020

[1] University of Modena and Reggio Emilia and CAPP, Center for the Analysis of Public Policies
Address: Viale Berengario 51, 41121, Modena, Italy
E-mail: michele.lalla@unimore.it

[2] University of Modena and Reggio Emilia and RECent, Center for Economic Research
Address: Viale Berengario 51, 41121, Modena, Italy
E-mail: maddalena.caviccioli@unimore.it

ISSN: 2281-440X online

Abstract

A (local) survey on income carried out in the city of Modena in 2002 generated four categories of units: interviewees, refusals, noncontacts, and sometimes unused reserves. In this study, all units were matched with their corresponding records in the Ministry of Finance 2002 databases for fiscal incomes of 2001 and the 2001 Census. Considering all four categories, participation increased by education level and activity status, while it decreased among low or high incomes. Considering interviewees only, over- and under-reporting, as well as measurement errors, were investigated by comparing the surveyed income with fiscal income. Age and level of income were the main covariates affecting the behaviours of taxpayers.

Keywords

Fiscal income, surveyed income, unit nonresponse, response bias, fiscal under-reporting

JEL codes

C46, C81, C83, D31

Acknowledgements

The authors are grateful to Professors Kapteyn and Ypma for sharing useful information regarding codes. Furthermore, the authors owe special thanks to Professor Massimo Baldini, who made surveyed income data available to us.

Index

1. Introduction

Sample selection and data collection have different sources of non-sampling errors (Lessler and Kalsbeek 1992). Frame errors (1) concern a lack of coincidence between the sampling frame and the target population. Nonresponse errors (2) involve a failure to obtain information on the units selected for the sample, concerning a single or a few questions (item nonresponses or missing values) or the entire questionnaire (unit nonresponses), due to refusal, untraceability of the unit, persons not being at home, incapacity to answer, and so on. Observation errors (3) refer to differences between recorded values in the data collection phase and their corresponding true values, supposing that the latter exist. Observation errors may be distinguished as measurement errors (involving the interviewer, the respondent, the questionnaire, and the interview mode) and processing errors (involving coding, transcription, imputation, editing, outlier treatment, and so on). Surveys are costly and the available funds, under time constraints, generally entail limits for reducing errors, as well as for the analysis thereof. In fact, every type of error has unpredictable consequences on estimates as the errors do not usually occur at random and their analysis in different situations could shed light on the characteristics of their effects.

Nonresponses are an intriguing issue and have long been recognised as a major problem in surveys (Hansen and Hurwitz 1946). Item nonresponses have been widely investigated, when they refer to units with a single or multiple missing items (Särndal and Lundström 2005; Brick 2013; Olson 2013), although the suggested procedures/models to remove/handle the missing items are often complex and unsatisfactory and the units involved are frequently left out from the standard analysis. Unit nonresponses are rarely processed and only statistical adjustment of survey weights is applied to reduce selection bias, given that nothing can be done as nothing is known about them. If some data on sampled units are available in administrative archives, then it is possible to combine corresponding observations from these datasets with those from the surveyed dataset, merging on one or more key variables. In the resulting joined dataset, all items of unit nonresponse interviews might be considered as item nonresponses (Durrant and Steele 2009; Petychev 2012), but this approach is difficult and suffers from the previously mentioned drawbacks. In fact, unbiased estimates of objective parameters are obtained only when all the units selected for the sample provide all the requested information, implying a response rate equal to 100%, which is practically impossible. Descriptions of the features of unit nonresponse errors in general, and specifically in income surveys, are a rarity (Korinek et al. 2006; Bollinger and Hirsch 2013).

Observation errors, called measurement errors without any kind of distinction, are dealt with in all of the literature on income surveys dating back to early inquiries of this type (Bancroft 1940) on through to more recent ones (Abowd and Stinson 2013; Paulus 2015); for a comprehensive review, see Bound et al. (2001), but also Alm (2012) and Pickhardt and Pinz (2014). A profile of measurement error is usually built by comparing survey data with matched administrative databases, as the latter are generally assumed to be reliable and error-free, although sometimes their definitions may differ from those of interest to the researcher. The error is obtained as the difference between the corresponding variables of the two sources, but it is likely that no database is error-free. Administrative databases are often constructed by linking data from several sources, which involves potential mismatches due to imperfections in merging information, such as the identification codes and joining variables. In the case of earnings, merging

Surveyed Income (SI) data with Fiscal Income (FI) data may provide (1) further suggestions about income measurement errors, (2) a rare profile of the corresponding effect of nonresponses on income (Jäntti 2004; Lalla et al. 2012), and (3) a possible estimation of tax evasion (one of the first studies is by Baldini et al. 2009), which is not dealt with in this paper for the sake of brevity. Thus the issue is the identification of the true measure between SI and FI. On the one hand, SI data are often assumed as true income in order to obtain a measure of fiscal under-reporting, but such data can contain many sources of biases and imprecisions (Moore et al. 2000; Schräpler 2006). On the other hand, FI data are obtained through an accurate examination of documents involving a precise amount for each individual, but they too can be distorted by erosion, elusion, and evasion (hereinafter all of which are referred to as evasion), and by potential mismatches when survey records are joined to fiscal records. Therefore, there are two unreliable manifest variables and the true income variable turns out to be a latent variable. Kapteyn and Ypma (2007) propose a mixture factor model for survey and register earnings data in which they relax the assumption that administrative data are error-free, which was the traditional operating standard in empirical applications, and allow for errors in the matching of the administrative data with the survey data (see also Meijer et al. 2012). These and other methods have been introduced to address the concern that the unit nonresponse rate is increasing and, consequently, the bias of estimates increases (among others, see Bee et al. 2015).

This paper focuses on the sources of errors in the context of an income survey.

Frame errors were avoided by setting up a good list of the target population units and applying a correct sampling technique to it. In fact, a statistically drawn sample is always representative of the population units reported in the list from which the sampling units were selected (Särndal et al. 1992).

Unit nonresponse errors in the income survey were investigated by comparing the administrative (fiscal) income data of responding units and nonresponding units. This type of survey is the first one ever conducted in Italy and a rarity in the literature. In fact, many validation studies deal only with item nonresponses and measurement errors in earnings survey data, but also in numerous other types of survey data collected through interviews, such as those that interview firms and consumers for industrial or marketing research (de Bruwer 1995; Collier and Bienstock 2007). Additionally, techniques used to handle item nonresponses have been developed in a number of studies (Little and Rubin 1987; Franses et al. 1999; Stocké 2006). However, studies on unit nonresponses are very rare because in a one-shot survey or in the first wave of a panel study, refusals and noncontacts are completely unknown, i.e., it is almost impossible to obtain information on them to gain an understanding of the mechanism inducing their behaviour or the factors determining their failure to participate. Therefore, analyses of unit nonresponse are prevailingly carried out on longitudinal survey data, starting at least from the second wave, because in the waves subsequent to the first, the data collected in the previous waves or the first wave are available concerning the individuals who refused the interview or were untraceable (Cannari and D'Alessio 1992; Campanelli and O'Muircheartaigh 1999).

Using individual fiscal codes, the records in the City Council of Modena's 2002 fiscal database for reference income year 2001 (containing gender, age, various types of income, etc.) were matched with the corresponding records in the 2001 Italian Census (containing gender, age, education level, occupation, etc.), and using the individual fiscal codes once again, they were also matched with each individual selected for the survey, to obtain four types of samples: interviewees, refusals, noncontacts, and unused reserves. Interviewees obviously refer to participating units. Refusals and noncontacts concern a usual classification of unit nonresponses (Brick 2013). Refusals indicate the group of units resulting from interviewer inability to persuade the contacted sampled units to respond or to persuade someone else, such as a gatekeeper, to provide access to a sample unit. Noncontacts regard inaccessible units, which may arise for a number of reasons: unit identification data may be incorrect or out of date, the individual may not be at home during the interview or call period, and/or the survey schedule may limit the number of contact attempts. The unused reserves constitute the group of households selected to replace refusals and noncontacts following a specified rule (see below), but they were not used for this aim because the selected households participated in the survey. The use of the fiscal code reduces mismatches to zero because it is an accurate and checked datum.

Measurement errors of income, i.e., the discrepancies between SI, which is the individual personal income collected through the survey, and FI, which is the individual income filed with the Italian Revenue Agency, may be analysed through this data set by comparing these two measurements at the individual level for interviewed taxpayers only.

The aims of the present paper, therefore, are the following.

The *first aim* concerns the unit nonresponse errors pursued by examining the differences in the participation rates and FI among the various samples (interviewees, refusals, noncontacts, and unused reserves) in the first wave of an income survey, with respect to the individual characteristics affecting both participation rates and the amount of FI.

The *second aim* deals with measurement errors made by each single individual as, at this stage, the units of the analysis were the interviewed individuals who are income earners and/or taxpayers as well. The differences between the SI and FI in the sample of interviewees were considered as errors and analysed with respect to the individual characteristics affecting both SI and FI. Although the results cannot be generalised to the whole country (Italy) due to the limited geographical area examined, they confirm empirical evidence reported in the literature and suggest interesting directions for further study. In fact, following a description of the general context, this study represents a rare opportunity to compare two measures of income obtained through fiscal and survey procedures.

The paper is organised as follows. Section 2 discusses some problems related to item nonresponses and measurement errors in the survey data and briefly reviews the main empirical findings. Section 3 describes the key characteristics of the survey and the fiscal data, as well as the basic features of the sample types: interviewees, refusals, noncontacts, and unused reserves. Section 4 illustrates, separately, the factors affecting unit nonresponses, refusals and noncontacts, and participation in the survey as well, using a multinomial logit model. It also analyses the factors affecting the behaviour of taxpayers and determining variations in FI, using ordinary multiple regression models, over the different types of samples: interviewees, refusals, noncontacts, unused reserves, and the pooled samples. Section 5 presents the determinants affecting over- and under-reporting in an income survey. Section 6 illustrates the estimation results of different measurement errors in responses obtained via the mixture factor model. Finally, Section 7 concludes with some comments and indications for future research.

2. Nonresponse and measurement errors in surveys on income

In almost all surveys there are units who do not respond, implying that the nonsampling errors are physiological traits of all concrete surveys, particularly those that collect income data. These errors may depend on a variety of factors and can affect inferences from survey data on income (Meyer and Mittag 2019; Burton et al. 2020).

Little is known about the effects of (unit) nonresponses.

Knowledge of the behaviours of non-respondents and the characteristics affecting them are often based on analyses usually carried out in waves subsequent to the first wave of panel surveys. Therefore, general conclusions are difficult to draw because effects and behaviours depend on the type of surveyed target variable and often on individual characteristics such as age, education level, and social status. Hence, there are many theories of survey participation. Each explains the behaviours of interviewees in a given sociological and/or psychological framework (Tourangeau et al. 2000) and provides strategies and suggestions to increase response rates looking at survey materials, techniques for interviewing, and characteristics of the sampling unit. Therefore, such theories focus on tool errors and when they consider the dynamics of interviewing, attention is drawn to the errors of the players. The latter may be reduced through interviewer training, which helps interviewers persuade selected sampling units to participate in a survey and collect high quality data (Couper and Groves 1992; Groves et al. 2002). However, training is not the whole story because the interviews are subject to restrictions concerning time and logistics, as well as to human and financial resource constraints, which affect the sampling design and strategies for data collection. Moreover, surveys on income, private property, and savings are burdensome for interviewees and involve the latter in various and specific sources of error (Curtin et al. 1989; Hurd et al. 2003). For example, the analyses of the data collected by the Survey of Household Income and Wealth (SHIW), carried out every two years by the Bank of Italy (Banca d'Italia 2004) has shown some empirical evidence of non-participation, that is, nonresponses were more frequent for households in urban areas and in the North, participation rates declined as income increased and household size decreased, the age of the head of the household affected the reliability of responses in a more complex way (Cannari and D'Alessio 1992), and unit nonresponses grew among wealthier households (D'Alessio and Faiella 2002). Data for the USA have revealed that response rates across states varied inversely with income, conditional upon other covariates (Korinek et al. 2007).

Response bias largely seems to be a fixed effect. The inclusion of imputed values replacing the missing data is not a solution. In fact, imputations are obtained from respondent data and are generally generated under the assumption that nonresponses are either random or ignorable (Bollinger and Hirsch 2013), which is empirically unsustainable, especially in income surveys. Through an understanding of the process leading from the sample selected to the sample observed, knowledge of the characteristics of non-respondents could be useful for increasing the quality of a survey and the representativeness of the resulting sample, and, perhaps, for quantifying the bias attributable to nonresponses. This approach will be adopted here to analyse the impact of individual characteristics on FI, which is a proxy of SI, with respect to various types of samples.

Measurement errors could partially overlap item nonresponse errors because the former could originate from ambiguous or inadequate statements. Classification of measurement errors is thus based on the generating cause: (1) instrument errors referable to the questionnaire, (2) technique errors deriving from the methodology and strategies used for data collection, (3) interviewer errors concerning causes or factors within the dynamics of the interview, and (4) errors of interviewees arising within the dynamics of the interview. In the last two cases, the errors depend on the personal characteristics of both figures (gender, age, education level, and personality), their comprehension or recollection of past events, whether they are qualified to answer and willing to be truthful, as well as the conditions created during the interview (see, among others, Tourangeau et al. 2000; Moore et al. 2000; Biancotti et al. 2004).

Measurement errors generally refer to the difference between the collected or measured data for a give variable and the corresponding true values, which are taken from a reliable administrative source. In the present case, the former is SI and the latter is FI. This investigation scheme applied was drawn from early (Bancroft 1940) and more recent inquiries concerning income (Hariri and Lassen 2017; Burton et al. 2020). Many experiments have been carried out over time to improve collection techniques using more original efficient strategies or new technologies and they have revealed new evidence of the nature and mechanisms of errors in surveys on income data (Burton et al. 2020). FI comes from an administrative data archive (tax data), which is an important source, but it also suffers from aspects concerning availability, evasion, elusion, and erosion (Alm 2012). Recently, some errors in these types of sources have been documented (Bollinger et al. 2018; Wilhelm 2019) and income statistics are often estimated using income survey data (Fiorio and D'Amuri 2005; Matsaganis and Flevotomou 2010).

Measurement errors usually have an unidentified structure and their analysis often implies assumptions that are not always verifiable in the real world. Statistical imputation for missing data also implies assumptions, which can introduce problematic biases (Bollinger and Hirsh 2013; Bollinger et al. 2019), but there are many advantages to eliminating missing data, invalid extreme values and other errors.

The analysis of data affected by measurement errors presents major obstacles when the errors have a specific and generally unknown pattern, while assumptions regarding these patterns are often suggested by accommodation rather than persuasion, simplification rather than representation. The starting point is the belief that there are data affected by errors. Then some way of quantifying the error size is sought. The more frequent approach consists in finding a corresponding data source that is believed to be free of errors. Thus, the differences between the data from the two sources are measures of the error size. Evaluation of the accuracy of SI has often been carried out by linking cross population surveys to administrative data (Bound et al. 2001) and comparing the SI with the income in the administrative earnings records, generally limited to the category of employees (among others see: Bollinger 1998; Bingley and Martinello 2017). In an explorative description of data, the category of retirees was included together with employees (see below). Moreover, income may also include other components of revenue, such as rental income from buildings and land or capital gains, which can be easily evaded or eluded. Despite the difficulties, SI and FI were accurately determined following procedures which made the two quantities comparable (Baldini et al. 2009). The empirical results of data analyses in the literature have yielded a negative correlation between measurement error and true earnings: mean-reverting (Bound and Krueger 1991).

3. Data sources

3.1. Survey on Economic and Social Conditions of households in Modena

The survey on economic and social conditions of households in Modena (SESC-MO1, first wave) was carried out in 2002 by the CAPP (Centre for Analyses of Public Policies) of the University of Modena's Department of Economics and it was based on two-stage cluster sampling stratified in accordance with the socio-healthcare districts, in which the municipalities were the primary sampling units and the households were the secondary sampling units (Baldini et al. 2004). To achieve the size of the target sample, three supplementary units were selected as reserves for each selected sampling unit (the household). If the first unit refused to be interviewed or was definitely considered to be untraceable, the interviewer would contact the next unit on the list of three reserves, and so on. The process could end either because one unit on the list was interviewed or because all four units were contacted and all refused to be interviewed or were untraceable. At the end of the survey, four groups of selected units were obtained: respondents or interviewees, refusals, noncontacts or untraceable units, and unused reserves. Hereinafter these groups are referred to as the variable "types of samples".

The number of households interviewed in Modena was 589 out of a target of 637 households, involving 1387 individuals. The total number of selected households in Modena was 637 times four (one sampling unit plus three reserves), equal to 2548 units.

The success rate of the SESC-MO1 was 33.4%, which is comparable with that of the SHIW (34.3%) in its cross-sectional component and with that of other similar surveys (Appendix A). For example, the response rates of the surveys carried out by the International Social Survey Program (ISSP) for the Federal Republic of Germany vary over time (1999-2005) and region (west, east), ranging from 34.7% to 51.8% (Hüfken 2010). In 1994, the European Community Household Panel (ECHP) presented response rates varying by country (Peracchi 2002): Luxemburg (40.7%), Germany (47.7%), Ireland (55.8%), Denmark (62.4%), Italy (90.7%), Spain (67.0%) and Greece (90.1%). However, responding to the ECHP was mandatory in Italy, thus resulting in very high participation rates.

The main survey information characterising ISCEMO1 are the following.

- ICESMO1 was a specialist income survey. Its sampling design and a the questionnaire were very similar to those used in the SHIW (Banca d'Italia 2004) and those in the ECHP, which ran from 1994 to 2001 in the EU-15.
- The questionnaires were administered by professional or trained interviewers.
- The statistical units of the sampling design were the households.
- The statistical units in the present data analyses are *always individuals*.
- The SI was the total taxable income at the personal level net of tax because it was the quantity comparable with the corresponding fiscal income item (see Appendix B).
- The reference year for SI was 2001 so as to be comparable with the reference year for FI, which was 2001.

3.2. Fiscal Data for the Different Types of Samples

The fiscal database of the Ministry of Economics and Finance is strictly protected by privacy policies that make it unusable for selecting a good sample and for matching their records with the corresponding surveyed records. Unexpectedly, the fiscal database of

Modena, i.e., of all taxpayers in the city, became available four years after the survey was conducted, allowing for exact matching of the sample unit records, using their fiscal identification numbers, briefly referred to as "fiscal codes" above and below, with the corresponding records in the 2002 fiscal database containing data for 2001. Moreover, the 2001 census database conducted by the Italian Institute of Statistics was accessible.

The availability of the census information and the income declared to Fiscal Authorities is a rare and precious opportunity for income surveys because FI could be analysed with respect to some available factors shedding light on the bias in income estimates for the entire population and on the different behaviours of those groups showing different attitudes towards participation in a survey. There are many problems involved in using administrative data and there are several techniques to solve them (among others see: Consolini and Donatiello, 2013; Jäntti and Törmälehto 2013; Jäntti et al. 2013). However, in this case exact and reliable matching was applied because the fiscal code is always checked and without errors, except for very rare negligible events in the handled context.

The 2002 fiscal database was matched with the 2001 census database and the resulting file was matched with all the sampling units included in the survey design. Four samples were thus obtained: interviewees, refusals, noncontacts or untraceable units, and unused reserves (Table 1).

The main aspects characterising the administrative data base and the matching carried out using the individual fiscal codes (Jäntti et al. 2013) are the following.

- The fiscal dataset consisted of administrative archives, created by the Internal Revenue Agency (IRA) to achieve its specific objectives or mission and consisting mainly in the nation's tax records, generally at the individual level.
- The available data for each individual came from the form he/she filed with the IRA: the TF730, TF-Unico, or TF770. For some details, see Appendix B.
- The FI was the total taxable income at the personal level net of tax and referring to 2001.
- The entire population was covered. However, complex situations in this context and the rules regulating them generated specific cases: there could be individuals who did not have to file a tax form, implying FI=0, but they may have declared an income in the interview (SI≠0) and vice versa individuals with FI≠0 and SI=0.

It should be noted again that the research was not planned a priori, as it was not possible to access the fiscal data set. However, unforeseeably it became available allowing the analysis of participation rates and measurement errors. The analysis suffers from this situation. For example, during contacts, typically, the area characteristics (such as local measures of deprivation, housing conditions, etc.) are observed. Here, not.

In Table 1, with respect to the sample of interviewees, the fiscal codes of family members were available for all 1387 subjects interviewed in Modena, but only 1098 of these subjects had information on SI and/or FI.

The bias arising from the data collection process, with respect to the participation of the selected individuals, may be examined in Table 1, considering the percentage difference in the mean FI for each cell (ij) and the corresponding marginal mean FI (i) in the column reporting the total: $\%\mathrm{D}_{ij} = 100 \cdot (\bar{y}_{\mathrm{FI};\,ij} - \bar{y}_{\mathrm{FI};\,i\cdot}) / \bar{y}_{\mathrm{FI};\,i\cdot}$. The lowest negative percentage difference with respect to the row total was observed for men (−8.0%) in the sample of noncontacts, followed by women in the sample of refusals (−6.7%), but neither was significant. In the sample of interviewees, men yielded the highest significant positive percentage difference (+12.3%, p=0.013), which is very large in the world of official statistics. All the interviewees showed a significant percentage difference of +8.4% (p=0.014). These figures represent relevant empirical findings and suggest that in income surveys one should expect overestimates on the average, i.e., there is a high probability of including individuals with incomes above average values.

Table 1. Descriptive statistics on individual fiscal income in 2002 by gender and by type of sample (TOS)

Gender\ TOS		**Interviewees**	**Refusals**	**Non-contacts**	**Unused reserves**	**Total**
Man	N	527	552	331	790	2200
	Mean	29479.0	25823.8	24138.0	25247.1	26238.6
	SD	29891.7	40663.5	28203.3	25068.5	31248.6
	%D	*12.3*	*–1.6*	*–8.0*	*–3.8*	
Woman	N	571	641	331	737	2280
	Mean	16696.6	15007.6	15916.5	16620.6	16083.9
	SD	13274.8	18350.3	15272.3	14474.6	15511.0
	%D	*3.8*	*–6.7*	*–1.0*	*3.3*	
Total	N	1098	1193	662	1527	4480
	Mean	22831.7	20012.2	20027.3	21083.5	21070.6
	SD	23681.5	31212.9	23032.2	21084.7	25052.0
	%D	*8.4*	*–5.0*	*–5.0*	*0.1*	
Missing		289	311	302	544	1446

Note: SD= Standard Deviation. %D= Percentage Difference of the mean and in each cell *ij* of this table, $\%\mathrm{D}_{ij} = 100 \cdot (\bar{y}_{\mathrm{FI};\, ij} - \bar{y}_{\mathrm{FI};\, i\cdot}) / \bar{y}_{\mathrm{FI};\, i\cdot}$. Moreover, note that 1446 is the number of individuals who had not filed a tax form in the reference period 2001, while N=5926 [=2200 (men) + 2280 (women) + 1446 (fiscal income missing)] is the total number of individuals involved in the 2548 selected households.

The distribution of FI was not the same for the four samples (Table 2). In fact, the comparison by means of a Kolmogorov-Smirnov test carried out in the six possible pairs of samples showed differences that were statistically significant at a 0.01 level for all pairs, except for the pair consisting of refusals and noncontacts (p=0.102). With respect to the total distribution, on the one hand the interviewees tended to under-represent FI lower than 17.5×10^3 and to over-represent FI greater than 17.5×10^3. On the other hand, refusals over-represented incomes lower than 17.5×10^3, but the percentage in the last class, $(100–500)\times10^3$, was somewhat high, implying that wealthy individuals tended to refuse to be interviewed about the economic status of their household more than other income levels. Noncontacts showed similar behaviour with less intensity than that of the refusals, but with a high percentage (11.3%) in the first class, i.e. $(\text{lowest} - 5)\times10^3$.

As is usual for income variables, the density distributions of FI showed long heavy right tails, especially for refusals, whose right tails reached €500,000, while the other samples reached values of about €300,000. However, in Figure 1, the maximum value reported on the abscissa is only €100,000 in order not to lose the shape of the histogram constituted by bars of equal width. For refusals, the median (€14,341) and the mode revealed the lowest FI values and the highest density values. For noncontacts the median value increased by 6.9%, for unused reserves it increased by 15.2%, and for interviewees it increased by 26.1%. In a similar manner, the modes increased and their corresponding density values decreased, as may be seen in Figure 1.

Table 2. Percentages and percentage densities of individual fiscal income (FI) in 2002, subdivided into classes, by type of sample (TOS)

FI\ TOS	Interviewees		Refusals		Noncontacts		UR		Total
classes	$f_{I\%j}$	$h_{I\%j}$	$f_{R\%j}$	$h_{R\%j}$	$f_{N\%j}$	$h_{N\%j}$	$f_{U\%j}$	$h_{U\%j}$	
Lowest – 5	8.8	1.766	8.5	1.694	11.3	2.266	10.6	2.122	9.7
05.0 – 10.0	13.3	2.660	21.8	4.358	17.4	3.474	15.0	3.000	16.7
10.0 – 15.0	16.8	3.352	24.0	4.794	20.1	4.018	17.3	3.458	19.4
15.0 – 17.5	8.7	3.496	10.9	4.360	11.5	4.592	10.5	4.216	10.3
17.5 – 20.0	9.8	3.936	7.0	2.816	7.6	3.020	8.8	3.512	8.4
20.0 – 22.5	9.5	3.788	5.0	1.980	5.7	2.296	8.1	3.248	7.3
22.5 – 25.0	5.8	2.332	4.4	1.744	5.4	2.176	6.0	2.384	5.4
25.0 – 30.0	8.7	1.748	6.2	1.240	5.6	1.118	7.1	1.428	7.1
30.0 – 40.0	7.2	0.719	4.4	0.444	6.7	0.665	6.9	0.688	6.3
40.0 – 60.0	5.8	0.292	4.1	0.206	4.5	0.227	4.7	0.236	4.8
60.0 – 100.0	3.7	0.093	2.3	0.057	3.0	0.076	3.6	0.090	3.2
100.0 – 500.	1.7	0.004	1.5	0.004	1.2	0.003	1.4	0.003	1.5
Total %	100.0	—	100.0	—	100.0	—	100.0	—	100.0
No. of cases	1098		1193		662		1527		4480

Note: $f_{k\%j}$ denotes the percentage distribution for k, indicating respectively: interviewees (I), refusals (R), noncontacts (N), and unused reserves (U). Similarly, $h_{k\%j}$ represents the corresponding class percentage densities for k indicating respectively: interviewees (I), refusals (R), noncontacts (N), and unused reserves (U, UR).

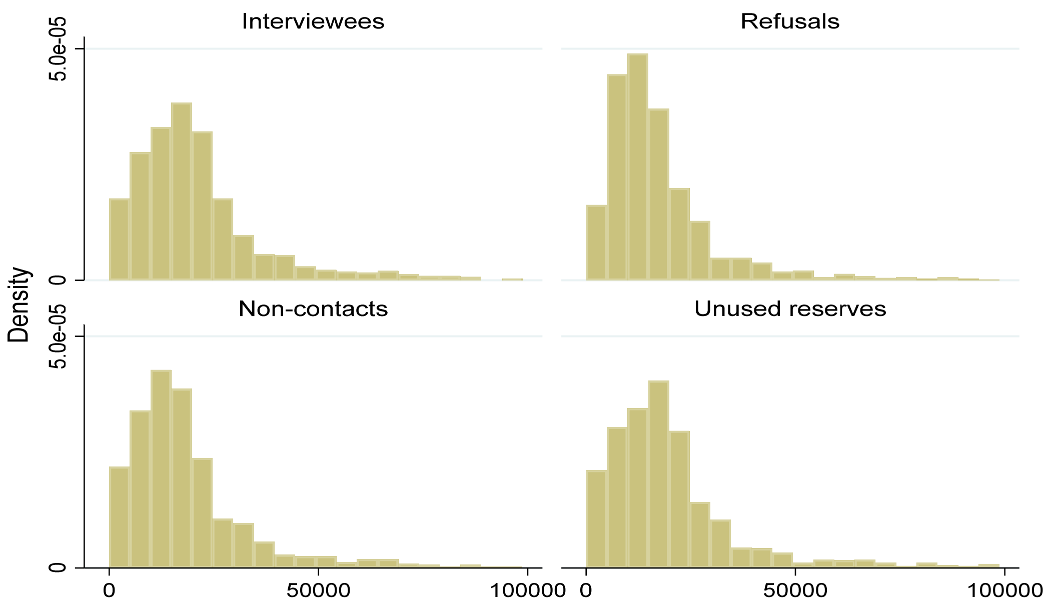

Figure 1. Histograms of Fiscal Income (FI) in 2001 by type of sample: 0<FI<100000; twenty bars of equal width=5×10^3€

4. Factors affecting unit nonresponses and selection bias

Even if not highly reliable and many values were missing, the plentiful set of auxiliary variables – obtained from matching with the 2001 Italian Census – could be used to explain the participation rate and the lack of participation (refusals or noncontacts), exploiting the rare occasion of having data on respondents and non-respondents at the individual level.

The unit of the analyses might logically appear to be the family. However, the characteristics determining participation in a survey are generally the attributes of an individual, such as gender, age, education, and so on, which do not refer to a group of individuals, as is a family. Therefore, the prominent limit concerned the decision-making individual as the sampling unit was the household and the decision to participate was made only by the person contacted by interviewer, and this person was often not the head of the household, but another cohabitant, usually a spouse. Individuals in the same household share many similar and correlated characteristics, which justifies the individual level of the analysis. In addition, as stated above, the two aims of the paper, unit nonresponse errors and measurement errors, were not the aims of the sampling design. They became possible aims and were identified four years after the survey was carried out. Therefore, the units of the present and subsequent analyses are always individuals. The factors affecting the FI in the different types of samples will now be discussed.

4.1. Multinomial logit for the data collection dynamics

At the first glance, the distinction between refusals and noncontacts in the variable type of sample (becoming the dependent variable here), may appear to offer no advantages, especially from the perspective of the analysis centred uniquely on the participation rate. However, the distinction is a usual one in data collection because the two groups may be different and there may be some implications for survey design, even if with respect to the modelling of the participation rate, they might have the same effects. Therefore, the distinction was maintained to ascertain the existence of this potential structural difference and the two groups were not pooled/ aggregated.

A bivariate analysis was carried out to identify factors affecting the dependent variable "type of sample" and the designated variables were FI, gender, age, education, activity status, occupational category, and sector of activity. The aggregation of the activity sector was based on NACE (Rev. 1.1 2002) and the terminology used for these variables is similar to that used by Eurostat (2009) in the EU-SILC (European Union – Statistics on Income and Living Conditions) and Atkinson and Marlier (2010): for details, see Appendix B. A multinomial logit model was then applied, assuming the polytomous variable "type of sample" as the dependent variable and the list of designated variables as independent variables, and fixing the unused reserves as the base alternative. Hence, the other possible outcomes (interviewees, refusals, noncontacts) were compared with the unused reserves. The Relative-Risk Ratios (RRR) were estimated and are reported in Table 3.

Interviewees decreased significantly for two factors: individual filing the TF-*Unico* (RRR=0.758, p=0.005) and individual filing the TF770 (RRR=0.749, p=0.007). The former was on the average wealthier than others, while the latter referred to the taxpayer who did not file a tax form and was poorer than others. Two significant factors increased the participation rate: upper secondary education (RRR=1.365, p=0.022) and tertiary education (RRR=1.812, p=0.001). In other words, the survey participation ratio increased with an increase in education level. Consequently, an increase in the estimation of income through the surveys could be expected, as high education levels generally corresponded to high levels of income. Moreover, richer and poorer individuals tended to have lower participation rates. This often involves an overestimation of income, given that wealthy people generally have high education levels. An emerging suggestion concerns survey organization, in the sense that skilful interviewers should be assigned to individuals who have low education levels or lower or higher incomes than others.

Table 3. Parameter estimates of a multinomial logit model for participation in a survey on income: base outcome is unused reserves

Type of sample	Interviewees			Refusals			Noncontacts		
Independent variables	RRR	SE	p<	RRR	SE	p<	RRR	SE	p<
TF-Unico	0.758	0.076	0.005	0.784	0.081	0.018	0.742	0.091	0.015
TF770	0.749	0.080	0.007	0.900	0.093	0.309	1.064	0.129	0.610
FI/100000	1.391	0.552	0.406	0.655	0.256	0.280	0.961	0.443	0.932
(FI/100000)^2	0.974	0.214	0.905	1.550	0.268	0.011	1.209	0.266	0.389
Women	1.141	0.100	0.134	1.029	0.092	0.747	1.012	0.105	0.907
Age/50	1.318	0.997	0.715	1.708	1.246	0.464	0.390	0.325	0.258
(Age/50)^2	1.034	0.381	0.928	1.122	0.389	0.740	2.004	0.817	0.088
EL: Primary (RG)									
EL: Lower secondary	0.987	0.130	0.920	0.722	0.088	0.008	0.714	0.108	0.026
EL: Upper secondary	1.365	0.186	0.022	0.767	0.102	0.046	0.870	0.143	0.397
EL: Tertiary	1.812	0.332	0.001	0.883	0.169	0.514	0.877	0.208	0.581
AS: Employed (RG)									
AS: Unemployed	0.717	0.252	0.343	0.897	0.329	0.767	0.809	0.271	0.526
AS: Retired	1.539	0.367	0.070	2.314	0.518	0.000	0.517	0.123	0.006
AS: Inactive	1.392	0.341	0.177	2.298	0.522	0.000	0.682	0.168	0.121
SL: Under-skilled (RG)									
SL: Low-skilled	1.024	0.183	0.896	1.468	0.270	0.037	0.867	0.180	0.492
SL: Medium-skilled	1.214	0.192	0.218	1.341	0.227	0.082	1.222	0.219	0.262
SL: High-skilled	0.908	0.201	0.662	0.689	0.178	0.150	0.869	0.243	0.614
SL: Manager	0.892	0.178	0.567	0.855	0.188	0.474	0.683	0.165	0.114
SA: Other Sectors (RG)									
SA: MEI of Section D	1.166	0.265	0.498	1.080	0.254	0.744	0.423	0.108	0.001
SA: Remaining D +C+E	0.731	0.177	0.195	1.121	0.266	0.629	0.796	0.190	0.339
SA: Trade & Transport	0.737	0.166	0.175	0.997	0.223	0.988	0.653	0.151	0.064
SA: Services	0.730	0.165	0.163	0.895	0.207	0.632	0.513	0.123	0.006
SA: PAEH	0.954	0.219	0.836	0.951	0.226	0.833	0.480	0.125	0.005
Constant	0.413	0.175	0.037	0.344	0.142	0.010	1.066	0.477	0.887

Note: RRR=Relative-Risk Ratios, SE= standard errors of RRR, PV= p-values, RG= Reference Group, EL=Education Level, AS=Activity Status, SL=Skill Level on the job, SA=Sector of Activity, MEI= Mechanical Engineering Industry of Section D (Manufacturing), PAEH= Public Administration + Education + Health.

Refusals decreased for three factors: TF-*Unico* referring to wealthy people (RRR=0.784, p=0.018), lower secondary education (RRR=0.722, p=0.008), and upper secondary education (RRR=0.767, p=0.046). There were also three factors that increased refusals: retirees (RRR=2.314, (p<0.001), other inactive persons (RRR=2.298, p<0.001), and persons with low skill levels on the job (RRR=1.468, p=0.037). Two opposite forces drove the behaviour of retirees and other inactive persons: enjoying conversation and interacting with other people, being overwhelmed by the fear of being deceived or robbed by strangers. Only the squared term of FI was significant (p=0.011), involving a slowly increasing probability of being part of the sample of refusals for increasing FI of up to about €80,000 and more accentuated decreasing probabilities for FI increasing to over €80,000. These results suggest that much attention should be devoted to elderly people often including retirees and other inactive persons.

Noncontacts revealed only significant factors that decreased the probability of being a "noncontact": TF-*Unico* referring to wealthy people (RRR=0.742, p=0.015), low education levels (RRR=0.714, p=0.026), retirees (RRR=0.517, p=0.006), and some specific sectors of activity (mechanical and engineering industry, services, and public administration plus education and health). As above, retirees were easily contacted as they are more likely to be at home. The squared term of age was significant only on a one-tailed test (RRR=2.004, p=0.088) involving a slowly increasing probability of being part of the sample of noncontacts for increasing ages up to about 52 years of age and more accentuated decreasing probabilities for ages over 52. There are various causes that generate noncontacts and it is difficult to prevent them. For example, it is well known that retirees and people with a limited education are very cautious in dealing with strangers and they do not open their door when someone rings the bell.

4.2. Determinants of fiscal income for respondents and non-respondents

To evaluate the effect on income levels deriving from the complex interdependence of the factors affecting the probability of participation in the survey, a regression model was estimated using the logarithm of FI as the dependent variable (regredend), as is usual in income data analysis because each estimated coefficient expresses the percentage change of the regredend for every unit that increases the corresponding independent variable, keeping the other independent variables constant. All the available variables were included in the models, which were estimated for each type of sample and for the combination of the four samples (full sample): interviewees, refusals, noncontacts, and unused reserves. For some remarks see Appendix B. The results are reported in Table 4.

The signs of the coefficients were consistent with those expected and described the empirical ascertained relationships between income and the explanatory factors, representing the structural differences in the distribution of income in the population. For example, women earn less than men (−28.8%), individuals with tertiary education level earn more than others, and retirees earn less than employees on the average. The expected relationship between earnings and age was an increase in earnings with an increase in age up to a specific age, generally around retirement, and thereafter earnings start to decrease with age. However, the age coefficients observed for the interviewees and for the full sample showed a substantial increasing trend. For example, in the sample of interviewees, the increase of FI with age turned around at 92.5 years: see Appendix B for details. The adjusted coefficient of determination (R-squared) was not very high, but it was acceptable, as is usually the case in models using prevailingly qualitative variables.

Tests for stability were carried out to ascertain whether there was a structural change among all the possible combinations of the various types of samples. The tests showed that the parameter values for interviewees were statistically different from those for refusals, noncontacts, and unused reserves, while the parameter values for the refusals, noncontacts, and unused reserves were statistically equivalent. The interesting and noticeable result that the occurrence of refusals and noncontacts generates a bias in the relationships between the determinants of income and its estimates requires caution when drawing conclusions, as discussed below.

The relevant outcomes concerned the differences between the coefficients in the sample of interviewees and the corresponding coefficients in the full sample, using the standard errors of the former in the test, because they involved the significant biases introduced by refusals and noncontacts. The factors significantly biasing the FI reported by the interviewees were upper secondary education level and self-employment, showing increases of 13.4% (p=0.045) and 21.5% (p=0.019), respectively. The other differences were not significant, but some are interesting all the same. For example, lower secondary and tertiary education levels and retirement did not show significant income increases in the sample of interviewees: 10.6%, 7.3%, and 4.9%, respectively. There were no factors significantly decreasing FI, with respect to the full sample. For example, the women interviewed showed a non-significant income decrease of 1.3% with respect to that of the full sample. Entrepreneurship and business partnership yielded high, but not significant income decreases: 21.0% and 14.4%, respectively. Note that these increases or decreases concern the sign of the differences between the two coefficients (those of interviewees minus those of the full sample) and not the impact of a single coefficient on the dependent variable. In the sample of interviewees, the effect of the age seemed slightly higher than that emerging in the full sample.

The model for the full sample allowed the variable "type of samples" to be introduced. Three dummy variables were included to distinguish the different samples. The sample of interviewees showed a borderline impact leading to an overestimated income of 6.2% on the average, which proved to be comparable with the total percentage difference (+8.4%) in Table 1. This finding is valuable because it is uncommon in the literature, but it is surprising at the same time, as it is opposite the results of other rare, similar studies. In fact, Cannari and D'Alessio (1992) obtained an underestimation of household income evaluated at 5.4% owing to non-participation (for SHIW 1987), analysing the non-response behaviour in the second wave of a panel sub-sample. D'Alessio and Faiella (2002) found an underestimation of 7% again. Bollinger and Hirsch (2013) estimated a 9% negative selection bias in responses among men. The peculiarity of our results concerned both the use of the first wave, instead of the second wave, and the availability of some useful data to analyse the behaviour of selected individuals. The previous analyses seem to indicate that selection bias operates in a different and complex way depending on various characteristics of the subjects. The behaviours of individuals classified as refusals or noncontacts in an income survey, with or without subsequent substitution of the unit nonresponses, may generate understatements or overstatements of income, but a positive bias of +6.2% was obtained here, also after having taken explanatory variables into account. Further investigations are necessary to verify the robustness of this finding.

Table 4. Parameter estimates of the regression model for the dependent variable ln(FI) and various samples

Type of sample	Interviewees		Refusals		Noncontacts		Unused reserves		All samples	
Independent variables	β	SE	β	SE	β	SE	β	SE	β	SE
Interviewees									0.062§§	0.035
Refusals									0.035	0.035
Noncontacts									0.058	0.041
TF-Unico	0.106	0.066	0.166*	0.076	–0.042	0.107	0.035	0.067	0.075*	0.038
TF770	–0.300**	0.063	–0.261**	0.065	–0.526**	0.087	–0.226**	0.062	–0.295**	0.034
Women	–0.288**	0.051	–0.315**	0.057	–0.264**	0.081	–0.246**	0.051	–0.275**	0.029
Age/50	2.116**	0.422	1.188*	0.472	1.398*	0.570	1.943**	0.395	1.726**	0.225
(Age/50)^2	–0.573**	0.201	–0.202	0.212	–0.122	0.267	–0.456*	0.189	–0.360**	0.106
EL: Primary (RG)										
EL: Lower secondary	0.247**	0.078	0.027	0.080	0.115	0.108	0.156*	0.074	0.142**	0.041
EL: Upper secondary	0.508**	0.079	0.224**	0.086	0.339**	0.108	0.400**	0.079	0.374**	0.043
EL: Tertiary	0.874**	0.098	0.770**	0.121	0.672**	0.161	0.788**	0.108	0.801**	0.058
AS: Employed (RG)										
AS: Unemployed	–1.148**	0.222	–1.259**	0.290	–0.846**	0.239	–1.320**	0.175	–1.133**	0.110
AS: Retired	–0.506**	0.132	–0.486**	0.169	–0.755**	0.179	–0.567**	0.131	–0.554**	0.073
AS: Inactive	–1.383**	0.134	–1.091**	0.177	–1.050**	0.181	–1.468**	0.130	–1.242**	0.075
ES: Employed (RG)										
ES: Entrepreneurship	–0.667*	0.277	–0.505	0.396	0.572	0.461	–0.647**	0.241	–0.457**	0.155
ES: Self-employed	–0.338**	0.104	–0.566**	0.137	–0.578**	0.161	–0.670**	0.094	–0.553**	0.058
ES: Partner	–0.581*	0.262	–0.429	0.280	–0.604*	0.288	–0.372*	0.165	–0.436**	0.112
SL: Under-skilled (RG)										
SL: Low-skilled	–0.048	0.115	–0.115	0.138	–0.024	0.153	0.085	0.095	0.011	0.059
SL: Medium-skilled	–0.063	0.101	0.034	0.126	0.041	0.131	0.122	0.087	0.053	0.053
SL: High-skilled	–0.054	0.131	0.186	0.196	0.201	0.208	0.100	0.128	0.055	0.077
SL: Manager	0.021	0.137	–0.168	0.178	–0.452*	0.197	–0.028	0.115	–0.097	0.073
Part-time	–0.379**	0.098	–0.440**	0.142	–0.657**	0.147	–0.287**	0.089	–0.400**	0.055
Temporary job	–0.278*	0.113	–0.177	0.152	–0.151	0.154	–0.210*	0.105	–0.215**	0.063
With paid workers	0.441*	0.185	0.481§§	0.274	–0.047	0.259	0.685**	0.148	0.466**	0.098
SA: Other Sectors (RG)										
SA: MEI of Section D	0.299*	0.132	0.502**	0.186	0.239	0.187	0.324**	0.126	0.399**	0.074
SA: Remaining D +C +E	0.197	0.145	0.431*	0.189	0.343*	0.169	0.269*	0.127	0.324**	0.076
SA: Trade & Transport	0.045**	0.141	0.411*	0.182	0.239	0.168	0.051	0.119	0.174*	0.072
SA: Services	0.232§§	0.140	0.353§§	0.185	0.145	0.176	0.137	0.121	0.248**	0.074
SA: PAEH	0.024	0.136	0.314§§	0.190	0.094	0.184	0.019	0.128	0.122	0.076
Constant	8.278**	0.243	8.762**	0.301	8.702**	0.315	8.292**	0.227	8.379**	0.132
Adjusted R-squared	0.439		0.307		0.324		0.378		0.364	
Residual Sum of Squares	627.5		948.2		497.0		1232.6		3381.6	
Number of cases, *n*	1098		1192		659		1524		4473	

Note: β= coefficients, SE= Standard Error, §§p<0.1, *p<0.05, **p<0.01. RG= Reference Group, EL= Education Level, AS= Activity Status, ES= Employment Status, SL= Skill Level on the job, SA= Sector of Activity, MEI= Mechanical Engineering Industry of Section D (Manufacturing), PAEH= Public Administration + Education + Health. The 7 cases having a negative FI were lost in the calculation of their logarithms.

5 Factors determining over- and under-reporting

The objective of the current analysis was to assess the determinants of measurement errors, as estimated through the difference in the percentages of SI and FI. Therefore, individuals who were not required to file an income tax return should be eliminated from the sample of interviewees. This was done by eliminating all individuals who had not filed a tax form (i.e., the entire tax form was missing) and had a SI equal to zero, i.e., individuals declaring a SI differing from zero (for them SI≠0 and FI=0) remained in the sample. Therefore, the unit of the analysis was still the *single individual with SI≠0 and/or FI≠0*, reducing the sample of interviewees to *n*=1031 (Table 5). It may be useful to examine the distributions of SI and FI reported in Figure 2. The two shapes were obviously very similar, even if the extreme values on the right and on the left had been had been eliminatd to obtain a readable graph, as in a high range of values, the tails tend to squash the central area. The distribution of FI shifted slightly on the right, revealing that on the average the values of FI were higher than the values of SI among the interviewees.

To illustrate the sample of interviewees, Table 5 reports ordinary descriptive statistics for SI, FI, and the percentage difference, %D, by employment status combined with job title and by gender. Percentage differences, $\%\mathrm{D}_i$, were calculated for each individual, *i*, assuming FI as the reference income, $\%\mathrm{D}_i = 100\left(\mathrm{SI}_i - \mathrm{FI}_i\right)/\mathrm{FI}_i$, notwithstanding it was affected by evasion, generating a contradiction in terms.

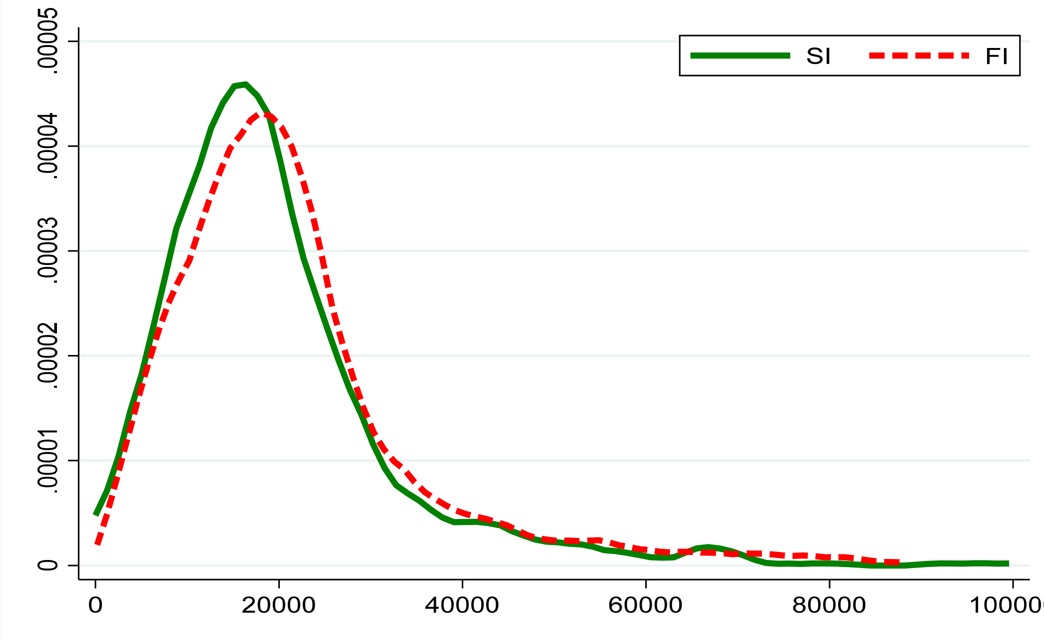

Figure 2. Epanechnikov kernel density estimates of Surveyed Income (SI) with 798 cases and Fiscal Income (FI) with 799 cases, both with reference year 2001

With respect to over- and under-reporting, the data in each cell *jk* of Table 5 may be interpreted according to at least two approaches: (1) considering the means of percentage differences, or (2) examining the differences of the means of SI and FI through their corresponding percentage differences. Symbols and formulae have been avoided for the sake of brevity, but note that the outcomes of the two approaches may differ and the comments reported below refer only to the means of the percentage differences that are statistically different from zero because the use of individual $\%\mathrm{D}_i$ is more coherent with the micro-modelling applied below.

Table 5. Descriptive statistics on Surveyed Income (SI), Fiscal Income (FI), and Percentage Difference (%D) by Employment Status (ES) combined with job title and by gender

Gender		Men			Women			Total		
ES & Job Title		N	Mean	SD	N	Mean	SD	N	Mean	SD
Entrepreneur	SI	15	27116	27591	12	16331	26614	27	22323	27192
	FI		47343	51628		13260	8319		32195	41981
	%D [a]		*–3.7*	*100.7*		*165.0*	*534.8*		*71.3*	*365.7*
Self-employed	SI	76	31287	34047	39	13937	25228	115	25403	32293
	FI		42939	46146		22454	20315		35992	40415
	%D [a]		*78.4*	*504.6*		*895.7*	*5433.4*		*355.6*	*3187.3*
Official-Executive	SI	61	50753	30499	18	31716	13005	79	46415	28582
	FI		54343	34880		36334	16120		50240	32408
	%D [a]		*–2.6*	*28.0*		*–9.9*	*15.4*		*–4.3*	*25.8*
Employee	SI	88	27352	17785	170	17653	7585	258	20961	12889
	FI		27695	20800		19161	7164		22072	14023
	%D [a]		*22.7*	*152.9*		*–5.4*	*28.7*		*4.2*	*92.9*
Labourer	SI	77	18221	5789	51	13402	6212	128	16301	6392
	FI		20423	5760		13478	6149		17656	6811
	%D [a]		*–9.9*	*20.1*		*104.5*	*688.5*		*35.7*	*435.9*
Unemployed	SI	3	0	0	9	2058	2265	12	1544	2144
	FI		2353	1960		4886	4384		4253	3999
	%D [a]		–100.0	0.0		–45.9	50.0		–59.4	49.2
Retiree	SI	156	19425	13182	185	12312	8513	341	15566	11447
	FI		21572	15376		15121	12378		18072	14180
	%D [a]		–3.9	28.9		6.2	179.8		1.6	133.8
Inactive	SI	19	124	540	52	6031	9846	71	4450	8811
	FI		3307	5177		10164	18805		8329	16549
	%D [a]		–97.2	12.2		–31.8	83.1		–49.3	76.9
Total	SI	495	25703	23645	536	14188	12068	1031	19717	19415
	FI		29764	29678		16797	13090		23023	23526
	%D [a]		8.5	211.9		75.0	1487.9		43.1	1082.8

[a] Note that the rows labelled with %D reports the mean and the standard deviation (SD) of the variable $\%\mathrm{D}_i = 100\,(\mathrm{SI}_i - \mathrm{FI}_i)/\mathrm{FI}_i$ for each individual *i*, that is $\overline{\%\mathrm{D}}$ and $\mathrm{SD}(\%\mathrm{D}_i)$. Therefore, these means of percentage differences differ from the percentage differences reported in Table 1. In fact, the latter were calculated using the means reported in Table 1.

On the average, women showed higher variability (heteroscedasticity) and higher over-reporting than men. Only the means of under-reporting percentage differences were significant for women: official executives (p=0.014), employees (p=0.014), unemployed (p=0.025), and inactive (p=0.008). Only the means of under-reporting percentage differences were significant for men too: labourer (p<0.001) and inactive (p<0.001). This approach presents some difficulties because the percentage difference has a lower bound equal to −100% and does not have an upper bound. In fact, Table 5 shows means of FI, $\overline{\mathrm{FI}}$, that are higher than the mean of SI, $\overline{\mathrm{SI}}$, with a positive mean of percentage differences, $\overline{\%\mathrm{D}}$. For example, self-employed men had $\overline{\mathrm{SI}} =$ € 31,287 and $\overline{\mathrm{FI}} =$ € 42,939, but $\overline{\%\mathrm{D}} = 78.4\%$. Although there were more under-reporters than over-reporters, this may happen because the latter may have much higher percentage differences than those of under-reporters and even greater than 100%. For other details and clarifications, see Appendix C.

Using the row of totals in Table 5, over-reporting on the average was $\overline{\%\mathrm{D}}$=8.5% for men and $\overline{\%\mathrm{D}}$=75.0% for women, and both were not statistically different from zero. Overall, non-significant over-reporting emerged from the data with a mean percentage difference of 43.1%, involving an apparent greater reliability of SI compared to FI, as is often assumed in practice (Fiorio and D'Amuri 2005; Matsaganis and Flevotomou 2010). However, this reliability is not a given, for as can be seen in Table 5, the total mean of SI (€ 19,717) is less than the total mean of FI (€ 23,023).

Given that FI was considered as the reference income, error was defined as the difference between SI and FI: $\Delta_i = \mathrm{SI}_i - \mathrm{FI}_i$. The latter were distinguished between positive and negative. Positive differences, (SI–FI)>0, constituted over-reporting and might represent measurement errors as well as evasion. Negative differences, (SI–FI)<0, constituted under-reporting and might represent measurement errors due to different causes, such as item nonresponses, inaccuracy, memory errors, and so on. The restricted sample of 806 units was used to investigate measurement error (Appendix C).

The distribution of the logarithm of errors, $\ln\left[\mathrm{abs}(\Delta_i)\right]$, showed a bell- shaped form (Figure 3), appearing to be mildly leptokurtic and negatively skewed for over- and under-reporters. Moreover, it can be noted that the number of under-reporters (61.8%) was higher than the number of over-reporters, involving a prevalent tendency among respondents to conceal their income in an interview (Moore et al. 2000; Schräpler 2006; Pickhardt and Pinz 2014).

The distribution of the logarithm of percentage difference of surveyed and fiscal income showed a bell-shaped form once again (Figure 4) and the same pattern. In the distribution there was also a truncation on the right deriving from the impossibility for under-reporters to attain a value lower than −100%, while it was possible for over-reporters to achieve a value greater than 100%.

The correlation between SI and FI, r(SI,FI), was equal to 0.810 for men and 0.708 for women, while the correlation between the measurement errors and the "true" values was $r(\Delta,\mathrm{FI})$=–0.468 for men and $r(\Delta,\mathrm{FI})$=–0.579 for women. The results observed for men are comparable with those reported by Bound and Krueger (1991), while the results observed for women are significantly higher than those reported by these authors. Therefore, mean-reverting was observed in these data too.

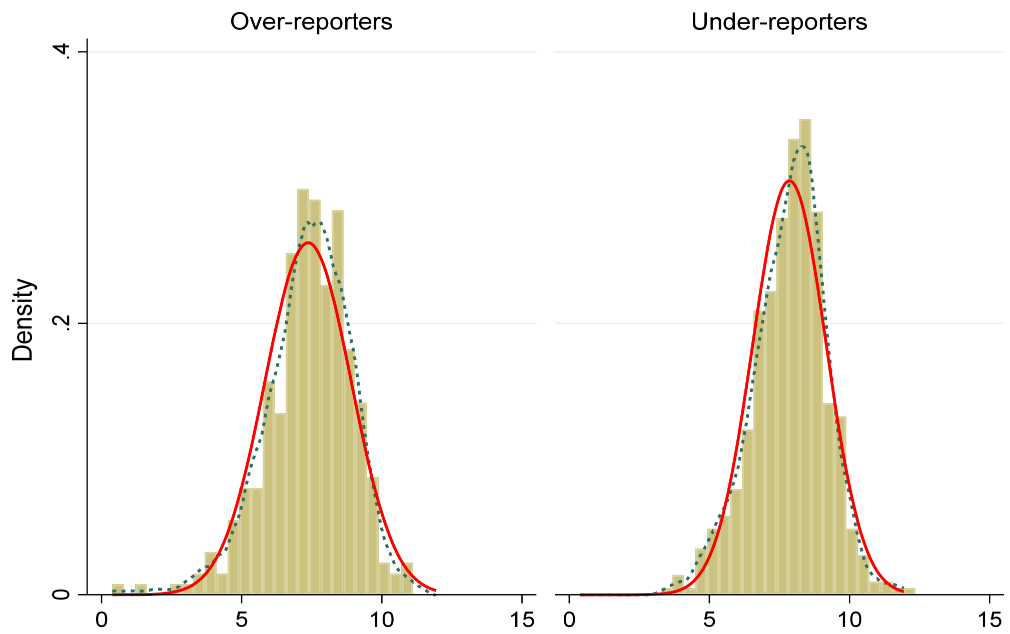

Figure 3. Epanechnikov kernel density estimates (dotted line) and corresponding normal density plot (solid line) of the logarithm of the absolute difference between Surveyed Income (SI) and Fiscal Income (FI) in 2001: (SI–FI)>0 indicates over-reporters (n=308) and (SI–FI)<0 indicates under-reporters (n=498)

Over- and under-reporting were analysed separately and the dependent variable of the model (regredend) was the logarithm of the absolute value of errors, $\ln[\text{abs}(\Delta_i)]$. To avoid potential dependence of the differences on the level of income, the percentage difference of SI and FI was considered as the regredend: $\ln\{\text{abs}[100\,(\text{SI}_i - \text{FI}_i)/\text{FI}_i]\}$. To explain the variability of measurement errors, a set of explanatory variables (regressors) was defined starting from data collected in the survey: type of tax form, FI, gender, age, education, employment status, sector of activity, tenure status of household (Eurostat 2009), degree of relationship among family members, and marital status.

The regressors already used in Section 4 entered the model with the same reference groups and modalities. The other modalities are listed in Table 6 and the details can be found in the Appendix C. The estimations of the models' coefficients are reported in Table 6.

For over-reporters, the percentage difference in errors was affected by several factors positively and negatively. Significant positive impacts were observed for the TF-*Unico*, implying that persons filing this form were evading more than others, but given the examined categories, their evasion presumably consisted of tax avoidance (elusion and erosion). Negative impacts were observed for retirees and people working in the public administration, education and health sectors. In fact, these categories are characterized by high tax compliance and the answers of these individuals were generally affected by response errors only. The squared term of income level had a positive coefficient, involving a negative and decreasing effect up to €57,098, given by $x_{\text{FI};T} = -(-9.187)/[2\times 8.045] \approx 0.57098$ multiplied by 100,000; thereafter the effect increased and became positive after about €114,000. A similar pattern emerged in the model with the regredend $\ln(+\Delta_i)$, except for the relationship with income level, in which only the coefficient of the squared term was significant, involving a quadratic increase in the observed range of fiscal income pertaining to over-reporters.

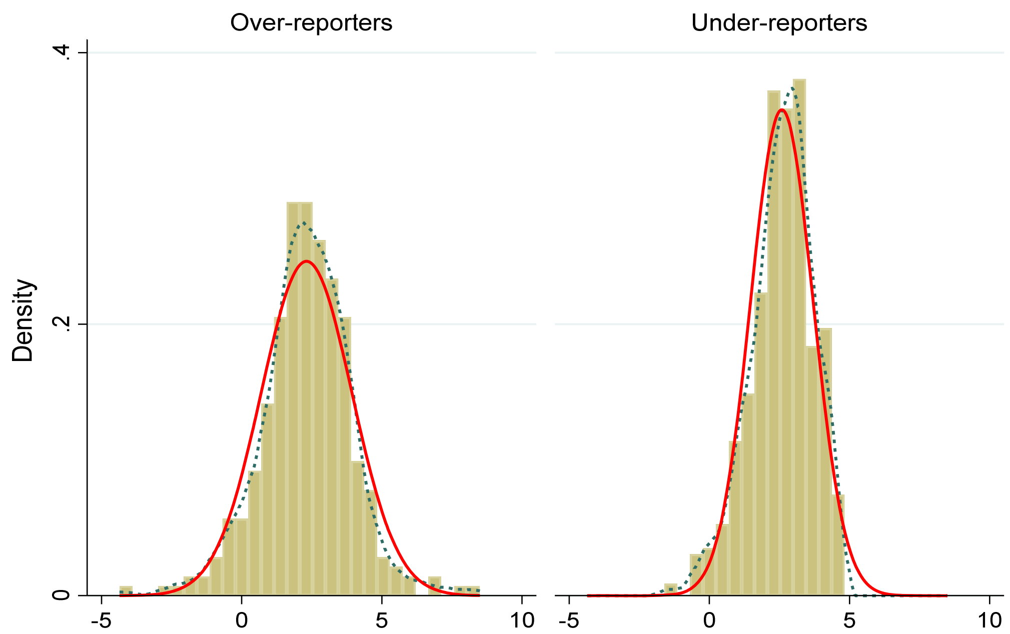

Figure 4. Epanechnikov kernel density estimates (dotted line) and corresponding normal density plot (solid line) of the logarithm of the absolute percentage difference of surveyed income (SI) and fiscal income (FI) in 2001, $\%D_i = 100\,(SI_i - FI_i)/FI_i$: $\%D_i > 0$ indicates over-reporters (n=308) and $\%D_i < 0$ indicates under-reporters (n=498).

For under-reporters, the percentage difference of errors was affected positively by the TF-*Unico*, women, and a post-verified cooperating attitude. Not only people filing the TF-*Unico* form, as above, but women also tended to be less tax compliant than other categories. The squared term of age was significantly positive involving a negative and decreasing effect on under-reporting up to 54.2 years of age, obtained as above multiplying by 50, after which the impact increased, but remained negative in any case. Negative coefficients were estimated for people with upper secondary and tertiary education levels, and for retirees. The coefficient of the squared term for income level had a negative coefficient involving an increasing effect on under-reporting up to €143,695, obtained as above, and thereafter it implied a decreasing effect, i.e., despite the fact that the percentage difference accounted for the level of FI, it increased nonlinearly as FI increased. A slightly different pattern emerged in the model with the regredend $\ln[\mathrm{abs}(\Delta_i)]$.

Table 6. Parameter estimates of the regression model for two different dependent variables and for over-reporters (+) and under-reporters (–)

Dependent variable	$\text{Ln}(+\%\text{D}_i)$†		$\text{Ln}[\text{abs}(-\%\text{D}_i)]$†		$\text{Ln}(+\Delta_i)$†		$\text{Ln}[\text{abs}(-\Delta_i)]$†	
Independent variables	Coefficient	SE	Coefficient	SE	Coefficient	SE	Coefficient	SE
TF–Unico	0.701**	0.246	0.438**	0.153	0.710**	0.226	0.434**	0.154
TF770	–0.041	0.244	–0.160	0.134	0.004	0.226	–0.251§§	0.134
FI/100000	–9.187**	1.878	2.598**	0.731			7.630**	0.733
(FI/100000)^2	8.045**	1.806	–0.904*	0.397	2.091	0.750	–2.737**	0.398
Women	–0.004	0.260	0.234§§	0.134	–0.014	0.239	0.156	0.134
Age/50	–2.159	2.186	–4.409**	1.115	–2.617	1.956	–3.921**	1.117
(Age/50)^2	1.081	0.986	2.033**	0.471	1.193	0.894	1.878**	0.472
EL: Primary (RG)								
EL: Lower secondary	–0.120	0.323	–0.184	0.150	–0.093	0.293	–0.108	0.151
EL: Upper secondary	0.377	0.347	–0.390*	0.181	0.406	0.300	–0.320§§	0.181
EL: Tertiary	0.257	0.386	–0.382§§	0.211	0.234	0.330	–0.309	0.212
ES: Labourer (RG)								
ES: Official, executive	0.362	0.465	–0.277	0.247	0.173	0.429	–0.374	0.247
ES: Employee	–0.030	0.316	–0.269	0.177	0.006	0.295	–0.213	0.178
ES: Retired	–1.479**	0.560	–0.576§§	0.305	–1.222*	0.523	–0.692*	0.306
SA: Other Sectors (RG)								
SA: MEI of Section D	–0.853§§	0.500	–0.291	0.272	–0.722	0.468	–0.240	0.273
SA: Remaining D +C +E	–0.713	0.519	–0.143	0.318	–0.450	0.487	–0.103	0.319
SA: Trade & Transport	–0.645	0.492	–0.228	0.299	–0.598	0.461	–0.239	0.300
SA: Services	–0.561	0.519	–0.153	0.328	–0.537	0.486	–0.150	0.329
SA: PAEH	–1.052*	0.455	–0.344	0.282	–0.830§§	0.426	–0.334	0.282
TSH: Others (RG)								
TSH: Tenant	–0.083	0.255	–0.101	0.143	–0.075	0.239	–0.055	0.144
TSH: Free	–0.066	0.369	0.004	0.212	–0.070	0.343	0.015	0.212
DFR: Others (RG)								
DFR: Partner	–0.032	0.309	0.141	0.150	–0.124	0.287	0.124	0.150
DFR: Daughter/son	–0.567	0.376	0.355	0.219	–0.498	0.353	0.326	0.220
MS: Others (RG)								
MS: Single	–0.293	0.302	–0.182	0.173	–0.303	0.283	–0.150	0.173
MS: Divorced	–0.009	0.449	–0.034	0.308	0.065	0.421	–0.104	0.309
MS: Widowed	0.346	0.400	0.009	0.192	0.407	0.372	0.100	0.192
Non-Cooperation for SI			2.045**	0.452			1.701**	0.453
Constant	5.474**	1.190	4.729**	0.714	9.207**	1.107	8.684**	0.715
Adjusted R-squared	0.130		0.125		0.152		0.362	
Number of cases, N	308		498		308		498	

Note: $\text{Ln}(\%\text{D}_i) = \text{Ln}[100\,(\text{SI}_i - \text{FI}_i)/\text{FI}_i]$ is the logarithm of percentage differences in surveyed income (SI) and fiscal income (FI): + indicates over-reporters (SI>FI) and – indicates under-reporters (SI<FI). $\text{Ln}(\Delta_i) = \text{Ln}(\text{SI}_i - \text{FI}_i)$ is the logarithm of the difference between the SI and FI.

§§p<0.1, *p<0.05, **p<0.01. RG= Reference Group, EL= Education Level, ES= Employment Status or activity status, SA= Sector of Activity, MEI= Mechanical Engineering Industry of Section D (Manufacturing), PAEH= Public Administration + Education + Health, TSH= Tenure Status of the Household, DFR= Degree of Relationship between Family members, MS= Marital Status.

6. Modelling different measurement errors

Recent developments (Kapteyn and Ypma 2007; Abowd and Stinson 2013) have showed that allowing for a richer error structure might shed light on potential biases and different sources of error in estimations. It is assumed that the true variable of a variable of interest (e.g. income) is not measured directly and both sources of data, administrative and survey files, may contain errors, although with a different structure. In particular, let ξ_i be the true value of the logarithm of income for an individual *i*, and let s_i and r_i be the values of income data collected by the survey and recorded in the administrative records, respectively. Firstly, define four independently and identically distributed and mutually independent normal variables: $\xi_i \sim N(\mu_\xi, \sigma_\xi^2)$, $\zeta_i \sim N(\mu_\zeta, \sigma_\zeta^2)$, $\eta_i \sim N(\mu_\eta, \sigma_\eta^2)$, and $\omega_i \sim N(\mu_\omega, \sigma_\omega^2)$, where i indexes the unit of observation. In the case of FI, the recorded value, r_i, is really unreliable as it is affected by evasion. Then, let π_r be the probability that the recorded FI is equal to the true FI of an individual *i*, ξ_i. Consequently, $(1-\pi_r)$ corresponds to the probability that the value r_i is lower than ξ_i, presumably involving a false declaration to the Fiscal Authorities. Let ζ_i be this false value. Kapteyn and Ypma (2007) assume that there is no correlation between ξ_i and ζ_i, but in our study, there was empirical evidence that evasion depended positively on income level, implying a correlation between ξ_i and ζ_i. Therefore, in an explorative perspective, the model is still interesting and the observed variable r_i is a mixture of correct matches and mismatches, specifically a mixture of two normal distributions:

$$r_i = \begin{cases} \xi_i & \text{with probability} \quad \pi_r \\ \zeta_i & \text{with probability} \quad (1-\pi_r) \end{cases}. \tag{1}$$

Let π_s be the probability that the surveyed value, s_i, is correct for the individual *i*. Thus, $(1-\pi_s)$ is the probability that s_i contains a response error, part of which is mean-reverting, as expressed by the term $\rho(\xi_i - \mu_\xi)$, where $\rho < 0$ implies a mean-reverting response error, as indicated by Bound and Kruger (1991), given that μ_ξ represents the mean of ξ_i. A proportion π_ω of these observations are contaminated, represented through an additional error term, ω_i. Contamination can result, for example, from processing errors, such as reporting errors (see above). Hence, the survey data, s_i, are a mixture of three different normal distributions with

$$s_i = \begin{cases} \xi_i & \text{with probability} \quad \pi_s \\ \xi_i + \rho(\xi_i - \mu_\xi) + \eta_i & \text{with probability} \quad (1-\pi_s)(1-\pi_\omega) \\ \xi_i + \rho(\xi_i - \mu_\xi) + \eta_i + \omega_i & \text{with probability} \quad (1-\pi_s)\pi_\omega \end{cases} \tag{2}$$

The distribution of s_i and r_i in (1) and (2), considered simultaneously, is a bivariate normal mixture, with 2×3 = 6 classes (Meijer et al. 2012). The estimation of the model's parameters and its derived moments is performed by the Maximum Likelihood method (further calculations and a detailed description of the estimation procedure can be found in Kapteyn and Ypma (2007), pp. 539-540). As noted in Kapteyn and Ypma (2007), it is quite remarkable that such a rich structure can be identified. However, this is a direct result of the non-normality of the error structure. Meijer and Ypma (2008) provide simple proof of identification for the case of a mixture of two normal distributions, of which the model (1)-(2) is a generalization.

The full model is then estimated together with three other constrained models: a model without contamination of the survey data (with π_ω=0), a model where no mismatching occurs (with π_r =1), and a model where both are left out, which is the basic model. Estimated values are reported in Table 7. The most interesting result is that allowing for mismatches or a contaminated sample in the model leads to a drop in the estimated value of the mean reversion parameter, ρ. Only when contamination and mismatches are not present do we observe substantial mean reversion, so that this phenomenon may not be as important as previously thought. This result is in line with Kapteyn and Ypma's findings with Swedish data.

Table 7. Parameter estimates of the mixture factor model using log earnings

	Full Model		No Contamination		No Mismatch		Basic Model	
	Coefficient	SD	Coefficient	SD	Coefficient	SD	Coefficient	SD
Log-lik.	–1702.81		–1734.50		–1782.56		–1824.67	
μ_ξ	9.801	.025	9.672	.042	9.107	.045	9.439	.051
σ_ξ	0.716	.022	1.032	.035	0.998	.044	1.325	.039
μ_ζ	9.454	.191	10.775	.498	–	–	–	–
σ_ζ	1.328	.156	0.899	.034	–	–	–	–
μ_ω	–3.329	.814	–	–	–1.702	.364	–	–
σ_ω	0.241	.459	–	–	0.887	.093	–	–
μ_η	–0.086	.015	0.471	.129	–0.379	.082	–	–
σ_η	0.266	.015	0.387	.013	0.399	.056	0.928	.056
π_r	0.921	.019	0.812	.027	–	–	–	–
π_s	0.287	.017	0.223	.030	0.194	.028	–	–
π_ω	0.046	.013	–	–	0.093	.018	–	–
ρ	–0.039	.027	–0.077	.041	–0.063	.040	–0.292	.053

The covariates are included through the parametrisation of μ_ξ as a function of individual characteristics. The "true" variable, ξ_i , is the dependent variable regressed on explanatory factors such as gender, age, and education:

$$\xi_i = \mathbf{x}_i \boldsymbol{\beta} + \varepsilon_i \tag{3}$$

The parameter estimates of the model, expressed by the previous three equations, are reported in Table 8. Here the covariates used are dummies for gender, age, age^2, education levels (lower secondary, upper secondary and tertiary), and retirement status. The conclusions that can be drawn are qualitatively similar to those found previously in this Section and in Section 5. In fact, gender, educational levels and the squared term of age are highly significant, confirming the importance of these covariates.

Furthermore, given the values of the estimated probabilities in the mixture distributions, we are able to compute the proportion of correct surveys, contamination and mismatches. Thus, the estimated percentage of correct survey data, π_s, was equal to 29%, while the fraction of contaminated survey values, $\pi_\omega(1-\pi_s)$, was between 3% and 7%. Finally, the proportion of mismatched administrative data, $(1-\pi_r)$, ranged between 8% and 11%, confirming that administrative databases are not always error-free.

Table 8. Parameter estimates of the mixture factor model using log earnings and covariates

	Full Model		No Contamination		No Mismatch		Basic Model	
	Coefficient	SD	Coefficient	SD	Coefficient	SD	Coefficient	SD
Log-lik.	–1686.62		–1704.88		–1772.09		–1793.65	
Women	–0.656	.011	–0.592	.043	–0.512	.033	–0.643	.018
Age	1.169	.340	0.987	.256	0.991	.204	1.134	.363
Age^2	–0.758	.002	–0.722	.004	–0.831	.008	–0.793	.004
LSE	0.199	.002	0.257	.006	0.272	.006	0.311	.009
USE	0.408	.038	0.366	.041	0.371	.043	0.498	.050
TE	0.597	.041	0.601	.053	0.624	.058	0.667	.059
Retired	–0.230	.025	–0.336	.029	–0.338	.032	–0.390	.040
μ_ξ	9.106	.107	9.568	.109	9.213	.107	9.883	.119
σ_ξ	0.865	.008	0.706	.006	0.884	.010	0.912	.023
μ_ζ	9.416	.128	10.212	.237	–	–	–	–
σ_ζ	1.137	.098	1.468	.113	–	–	–	–
μ_ω	–0.197	.078	–	–	–0.002	.001	–	–
σ_ω	1.002	.193	–	–	1.154	.223	–	–
μ_η	–0.220	.031	–0.365	.044	–0.277	.038	–	–
σ_η	0.158	.004	0.145	.003	0.167	.007	–	–
π_r	0.887	.015	0.841	.019	–	–	–	–
π_s	0.298	.017	0.299	.018	0.237	.014	–	–
π_ω	0.097	.008	–	–	0.164	.011	–	–
ρ	–0.025	.010	–0.022	.013	–0.038	.025	–0.245	.041

Note: *Log-lik.*= Log-likelihood. LSE= Lower Secondary Education level. USE= Upper Secondary Education level. TE= Tertiary Education level.

7. Conclusions

The unit nonresponses examined here refer to the first wave, which is a rarity in studies of this type. Despite the spatial limitation of the study, the results are in line with other results reported in the literature, but often carried out on the subsequent waves (D'Alessio and Faiella 2002; Korinek et al. 2007; Bound and Kruger 1991). In fact, the non-randomness of unit nonresponses was empirically confirmed, which undermines any strategies for replacement of non-respondents. Participation rates appear to increase among individuals with high education levels, and retirees, while they decrease among taxpayers placed at the extremes of the income distribution, i.e. wealthier and poorer individuals.

Refusals increased among retirees, other inactive persons, low and medium skill levels, and according to a parabolic trend with FI, implying that they tend to be high for low and high FI. Refusals decreased among individuals with lower or upper secondary education levels or for wealthier taxpayers.

Noncontacts increased according to a parabolic trend with age, implying that they are high among young and elderly people. Noncontacts decreased among wealthier taxpayers, individuals with a lower secondary education level, retirees, and some types of activity sectors: the mechanical engineering industry, trade and transport, services, public administration, education, and the health sector.

The FI bias observed among interviewees was +12.3% for men and +3.8% (not significant) for women, with respect to the expected means, involving a tendency to collect men and women who are wealthier than the corresponding means derived from a random process without disturbances. This is a very large number and indicates an overestimation of income in the survey. The interviewees showed an increase of 6.2% in the estimation of FI, keeping the available covariates constant, which is in contrast with other findings (Cannari and D'Alessio 1992; D'Alessio and Faiella 2002; Bollinger and Hirsch 2013).

The analyses of measurement errors showed that under-reporters were more numerous (61.8%) than over-reporters and with a higher mean and a more concentrated distribution than those of over-reporters, indicating a prevailing tendency of respondents to conceal their income, which is common to almost all people in any country. Many factors affected the two types of errors, but FI and age in a nonlinear form, retirees, and education level were the prevailing factors.

Furthermore, when we allow for the fact that both administrative and survey files may contain errors, using mixture factor models, we can observe a weakening of the mean-reversion hypothesis due to the presence of a richer error structure. In fact, the proportion of mismatched administrative values (around 10%) confirms that fiscal databases are not always error-free.

The main challenge for a generalization of the results in this paper for Italy or other wealthy countries regards data collection: unit nonresponses in a first wave panel study are usually completely unknown, making it impossible to detect the mechanism of nonparticipation. The results of this study are also completed by the rare opportunity to compare two measures of income in fiscal and survey procedures. In light of what has been discussed thus far, it follows that our techniques can be applied to other datasets insofar as they are available to researchers. Thus, generalization of results may be possible in the near future for other countries and further applications could also encompass estimations of the proportion of tax evaders.

References

Abowd, J. M., & Stinson, M. H. (2013) Estimating measurement error in annual job earnings: A comparison of survey and administrative data. *Review of Economics and Statistics*, **95** (5), 1451–1467.

Alm, J. (2012) Measuring, explaining, and controlling tax evasion: lessons from theory, experiments, and field studies. *International Tax and Public Finance*, **19**(1), 54–77.

Atkinson, A. B., & Marlier, E. (eds.) (2010) *Income and living condition in Europe*. Luxembourg: cat. No KS-31-10-555-EN-C, Publication Office of the European Union.

Baldini, M., Bosi, P., & Silvestri, P. (eds) (2004) *La ricchezza dell'equità. Distribuzione del reddito e condizioni di vita in un'area a elevato benessere*. Bologna: il Mulino.

Baldini, M., Bosi, P., & Lalla, M. (2009) Tax evasion and misreporting in income tax returns and household income surveys. *Politica Economica*, a. **XXV**(3), 333–348.

Banca d'Italia (2004) *I bilanci delle famiglie italiane nell'anno 2002*, Supplementi al bollettino statistico, anno XIV, no. **12**. Roma: Banca d'Italia.

Bancroft, G. (1940) Consistency of Information from Records and Interview. *Journal of the American Statistical Association*, **35**(210), Part 1, 377–381.

Bee, C. A., Gathright, G. M. R., & Meyer, B. D. (2015). Bias from Unit Non-Response in the Measurement of Income in Household Surveys. Available on the site accessed 19 March 2020: https://harris.uchicago.edu/files/jsm2015_bgm_unit_non-response_in_cps.pdf

Biancotti, C., D'Alessio, G., & Neri, A. (2004) Errori di misura nell'indagine sui bilanci delle famiglie italiane. *Temi di discussione*, no. **520**. Roma: Banca d'Italia.

Bingley, P., & Martinello, A. (2017). Measurement error in income and schooling and the bias of linear estimators. Journal of Labor Economics, **35**(4), 1117–1148.

Bollinger, C. R. (1998). Measurement error in the current population survey: A nonparametric look. Journal of Labor Economics, **16**(3), 576–594.

Bollinger, C. R., & Hirsch, B. T. (2013) Is earnings nonresponse ignorable? *Review of Economics and Statistics*, **95**(2), 407–416.

Bollinger, C. R., Hirsch, B. T., Hokayem, C. M., & Ziliak, J. P. (2018). The Good, The Bad and The Ugly: Measurement Error, Non-response and Administrative Mismatch in the CPS. https://www.semanticscholar.org/paper/The-Good-%2C-The-Bad-and-The-Ugly-%3A-Measurement-Error-Bollinger-Hirsch/4f183466ef6566cbec9b579aed0ce0e66c53ee55

Bollinger, C. R., Hirsch, B. T., Hokayem, C. M., & Ziliak, J. P. (2019). Trouble in the Tails? What We Know about Earnings Nonresponse 30 Years after Lillard, Smith, and Welch. *Journal of Political Economy*, **127**(5), 2143-2185.

Bound, J., & Kruger, A. B. (1991) The Extent of Measurement Error in Longitudinal Earnings Data: Do Two Wrongs Make a Right? *Journal of Labor Economics*, **9**(1), 1–24.

Bound, J., Brown, C., & Mathiowetz, N. (2001) Measurement Error in Survey Data. In J. J. Heckman and E. Leamer (Eds.), *Handbook of Econometrics* (vol. **5**, pp. 3705–3843). Amsterdam: Elsevier.

Brick, J. M. (2013) Unit Nonresponse and Weighting Adjustments: A Critical Review. *Journal of Official Statistics*, **29**(3), 329-353.

Burton, J., Crossley, T. F., Fisher, P., Gaia, A., and Jäkle, A. (2020). Understanding and reducing income reporting error in household survey. *Understanding Society Working Paper*, No. **2020–01**. Essex: Institute for Social and Economic Research (ISER), University of Essex. https://www.iser.essex.ac.uk/research/publications/working-papers/understanding-society/2020-01 Accessed 16 March 2020.

Campanelli, P., & O'Muircheartaigh, C. (1999) Interviewers, Interviewer Continuity, and Panel Survey Nonresponse. *Quality & Quantity*, **33**(1), 59–76.

Cannari, L., & D'Alessio, G. (1992) Mancate interviste e distorsione degli stimatori. *Temi di discussione*, no. **172**. Roma: Banca d'Italia.

Collier, J. E., & Bienstock, C. C. (2007) An analysis of how nonresponse error is assessed in academic marketing research. *Marketing Theory*, **7**(2), 163–183.
Consolini, P., & Donatiello, G. (2013). 9. Improvements of data quality through the combined use of survey and administrative sources and micro simulation model. In M. Jäntti, V.-M. Törmäletho, & E. Marlier, *The use of registers in the context of EU-SILC: challenges and opportunities*. Collection: Statistical working papers (pp. 125–71). Luxembourg: Cat. No: KS-TC-13-004-EN-N, Publication Office of the European Union. https://ec.europa.eu/eurostat/documents/3888793/5856365/KS-TC-13-004-EN.PDF Accessed 16 March 2020.
Couper, M. P., & Groves, R. M. (1992) The Role of Interviewer in Survey Participation. *Survey Methodology*, **18**, 263–278.
Curtin, R. T., Juster, F. T., & Morgan, J. N. (1989) Survey Estimates of Wealth: An assessment of Quality. In R. E. Lipsey and H. S. Tice (Eds.), *The Measurement of Saving, Investment, and Wealth*, National Bureau of Economic Research, Studies in Income and Wealth (vol. **52**, pp. 473–548). Chicago, IL: The University of Chicago Press.
D'Alessio, G., & Faiella, I. (2002) Non-response Behaviour in the Bank of Italy's Survey of Household Income and Wealth. *Temi di discussione*, no. **462**. Roma: Banca d'Italia.
de Bruwer, W. J. (1995) A Sampling Methodology and Replacement Procedure for Reducing Nonresponse Bias in Fixed-Address Industrial Surveys. *Journal for Studies in Economics and Econometrics*, **19**(2), 61–75.
Durrant, G. B., & Steele, F. (2009) Multilevel modelling of refusal and non-contact in household surveys: evidence of six UK Government surveys. *Journal of the Royal Statistical Society*, Series **A, 172**, Part 2, 361–381.
Fiorio, C. V., & D'Amuri, F. (2005) Workers' tax evasion in Italy. *Giornale degli economisti e Annali di economia*, **64**(2/3), 247-270.
Eurostat (2009) *Description of Target Variables: Cross-Sectional and Longitudinal*, Document EU-SILC 065/2009, Directorate F, Unit F-3. Luxembourg: Eurostat.
Franses, P. H., Geluk, I., & Van Homelen, P. (1999) Modeling Item Nonresponse in Questionnaires. *Quality & Quantity*, **33**(2), 203–213.
Groves, R. M., Dillman, D. A., Eltinge, J. L., & Little, R. J. A. (eds) (2002) *Survey Nonresponse*, New York: John Wiley & Sons.
Hansen, M. H., & Hurwitz, W. N. (1946) The Problem of Non-Response in Sample Surveys. *Journal of the American Statistical Association*, **41**(236), 517–529.
Hariri, J. G., & Lassen, D. D. (2017). Income and outcomes social desirability bias distorts measurements of the relationship between income and political behavior. *Public Opinion Quarterly*, **81**(2), 564–576.
Hüfken, V. (2010) Supplementary questionnaire and nonresponse-results from the German ISSP survey. *Quality & Quantity*, **44**(4), 607–622.
Hurd, M., Juster, F. T., & Smith, J. P. (2003) Enhancing the Quality of Data on Income. *Journal of Human Resources*, **38**(3), 758–772.
Jäntti, M. (2004) The effect of measurement errors, non-response, and attrition on income inequality, poverty and mobility. In M. Ehling and U. Rendtel (Eds.) *Harmonisation of panel surveys and data quality* (pp. 89–116). Wiesbaden: Statistisches Bundesamt.
Jäntti, M., & Törmäletho, V.-M. (2013). 4. Survey- and register-based estimates of income distribution and poverty. In M. Jäntti, V.-M. Törmäletho, & E. Marlier, *The use of registers in the context of EU-SILC: challenges and opportunities*. Collection: Statistical working papers (pp. 57–71). Luxembourg: Cat. No: KS-TC-13-004-EN-N, Publication Office of the European Union. https://ec.europa.eu/eurostat/documents/3888793/5856365/KS-TC-13-004-EN.PDF Accessed 16 March 2020.
Jäntti, M., Törmäletho, V.-M., & Marlier, E. (2013). *The use of registers in the context of EU-SILC: challenges and opportunities*. Theme 3. Collection: Statistical working papers. Luxembourg: Cat. No: KS-TC-13-004-EN-N, Publication Office of the European Union.

https://ec.europa.eu/eurostat/documents/3888793/5856365/KS-TC-13-004-EN.PDF Accessed 16 March 2020.
Kapteyn, A., & Ypma, J. Y. (2007) Measurement error and misclassification: A comparison of survey and administrative data. *Journal of Labor Economics*, **25**(3), 513–551.
Korinek, A., Mistiaen, J. A., & Ravallion, M. (2006) Survey nonresponse and the distribution of income. *Journal of Economic Inequality*, **4**(1), 33–55.
Korinek, A., Mistiaen, J. A., & Ravallion, M. (2007) An econometric method of correcting for unit nonresponse bias in surveys. *Journal of Econometrics*, **136**(1), 213–235.
Lalla M., Ferrari, D., & Frederic, P. (2012) Unit Nonresponse Errors in Income Surveys: A Case Study. *Quality & Quantity*, **46**(6), 1769–1794.
Lessler, J. T., & Kalsbeek, W.D. (1992) *Nonsampling Errors in Surveys*. New York: John Wiley & Sons.
Little, R. J. A., & Rubin, D. B. (1987) *Statistical Analysis with Missing Data*, John Wiley & Sons, New York.
Matsaganis, M., & Flevotomou, M. (2010). Distributional Implications of Tax Evasion in Greece. *Hellenic Observatory Papers on Greece and Southeast Europe GreeSE Paper No.* ***31***, The Hellenic Observatory, LSE (London School of Economics).
Meijer, E., & Ypma, J. (2008) A simple identification proof for a mixture of two univariate normal distributions. *Journal of Classification*, **25**, 113-123.
Meijer, E., Rohwedder, S., & Wansbeek, T. (2012) Measurement error in earnings data: Using a mixture model approach to combine survey and register data. *Journal of Business & Economic Statistics*, **30**(2), 191–201.
Meyer, B. D., & Mittag, N. (2019). Using Linked Survey and Administrative Data to Better Measure Income: Implications for Poverty, Program Effectiveness, and Holes in the Safety Net. *American Economic Journal: Applied Economics*, **11**(2), 176–204.
Moore, J. C., Stinson, L. L., & Welniak, E. J. Jr. (2000) Income measurement error in surveys: a review. *Journal of Official Statistics*, **16**(4), 331–361.
NACE, Rev. 1.1 (2002) http://ec.europa.eu/environment/emas/pdf/general/nacecodes_en.pdf. Accessed 24 May 2016.
Olson, K. (2013) Do non-response follow-ups improve or reduce data quality?: a review of existing literature. *Journal of the Royal Statistical Society*, Series **A, 176**, Part 1, 129–145.
Peracchi, F. (2002) The European Community Household Panel: A Review. *Empirical Economics*, **27**(1), 63–90.
Petychev, A. (2012) Multiple imputation for unit nonresponse and measurement error. *Public Opinion Quarterly*, **76**(2), 214–237.
Pickhardt, M., & Prinz, A. (2014) Behavioral dynamics of tax evasion – a survey. *Journal of Economic Psychology*, **40**, 1–19.
Paulus, A. (2015) Tax evasion and measurement error: An econometric analysis of survey data linked with tax records. *ISER Working Paper Series*, No. **2015–10**. Essex: Institute for Social and Economic Research, University of Essex.
Särndal, C.-E., Swensson, B., & Wretman, J. (1992) *Model Assisted Survey Sampling*. Berlin: Springer-Verlag.
Särndal, C.-E., & Lundström, S. (2005) *Estimation in Surveys with Nonresponse*. New York: John Wiley & Sons.
Schräpler, J. P. (2006) Explaining Income Nonresponse – A Case Study by means of the British Household Panel Study (BHPS). *Quality & Quantity*, **40**(6), 1013–1036.
Stocké, V. (2006) Attitudes Toward Surveys, Attitude Accessibility and the Effect on Respondents' Susceptibility to Nonresponse. *Quality & Quantity*, **40**(2), 259–288.
Tourangeau, R., Rips, L. J., & Rasinki, K. (2000) *The Psychology of Survey Response*. Cambridge: Cambridge University Press.
Wilhelm, D. (2019). Testing for the Presence of Measurement Error. *CeMMAP Working Paper* CWP 48/19. Old version: CWP 45/18.

Appendices

Appendix A. *SESC-MO1 versus SHIW and other details*

The survey on economic and social conditions of households in Modena (SESC-MO1, first wave) is structurally similar to the Survey of Household Income and Wealth (SHIW), which is carried out every two years by the Bank of Italy. It is also similar to the survey carried out by Istat as part of EU-SILC (European Union – Statistics on Income and Living Conditions). SESC-MO1 was comparable with one of the latter, for example SHIW, to evaluate the performance of the interviewers in data collection considering the data reported in Table A1, which shows the percentages of interviewees, refusals, noncontacts, and unused reserves, as well the corresponding data obtained in the SHIW (Banca d'Italia 2004).

Table A1 Absolute Frequencies (N) and Percentages of Households in the SESC-MO1 and SHIW by Type of Sample

Type of sample		Interviewees	Refusals	Non-contacts	Total units used	Unused reserves	Total
SESC-MO1 2002	N	589	704	472	1765	783	2548
	%	33.4	39.9	26.7	100.0	30.7	
SHIW 2002	N	8011	14179	1166	23356		
	%	34.3	60.7	5.0	100.0		

Note: SESC-MO1=Survey on Economic and Social Conditions of Households in Modena, first wave. The ineligible units were included in the noncontact sample. SHIW=Survey of Household Income and Wealth (SHIW) in Italy, carried out by the Bank of Italy. Non-contacts in SHIW (n=1166) do not include 476 (2.0%) non-existent households (tax register address no longer valid due to death, change of address or incorrect address), classified as ineligible units.

The SESC-MO1 used a questionnaire requesting information on many variables and specifically concerning net earnings, real estate, capital and financial assets over the last twelve months, thereby reducing the length of the recall period. Therefore, the interview period started in June 2002 because in that period people are more familiar with the data being requested, as June and July are near the deadline for filing tax forms. However, a few cases involved interviews as late as December 2002. Thus, the beginning of the reference income periods varied from June 2001 to December 2001 because the survey interval became a bit broader with respect to the target period, which was June-July or at most June-September, excluding August as it is a popular vacation month. As a consequence, the Surveyed Income (SI) reference period did not coincide with the calendar year, which was the Fiscal Income (FI) reference period and the SI comparable with FI was an estimated quantity. For this study, it would have been better to collect income over an annual period coinciding with the calendar year, but the opportunity to use fiscal information arose unexpectedly. In fact, the study was designed to optimise other aspects of the interviews and the displacements of the SI reference periods involved hypothetical and nontrivial procedures to obtain individual yearly SI to be compared with

FI. However, a detailed discussion of these operations and results is beyond the scope of this paper. Presumably, harmonisation procedures smoothed out the measurement errors, but also increased them through an additional estimated error.

As stated in the text of the paper, the 2002 fiscal database of the Ministry of Finance containing data for 2001 was matched with the 2001 census database and the resulting database was matched with the selected individuals, obtaining the final matched file. The procedures carried out on the final matched file and the results obtained are similar to a post-hoc analysis, given the steps followed in the data acquisition process. The overall sample included 6010 individuals, but for 84 of the latter there was no information available in the fiscal database, thus reducing the overall sample size to 5926. The 84 individuals were new-born babies in 2001 and without a fiscal code or new residents or foreigners still having their fiscal residence in another town or country. Therefore, the FI was not available and they were not eligible for the present study because each individual should have at least one of the two variables considered: SI and FI. Moreover, considering the objective of the analysis, individuals who were not required to file an income tax return were eliminated from the four samples (interviewees, refusals, noncontacts, and unused reserves): 1530 units (Table 1).

The interviewees constituted the sample used for the evaluation of errors made by individuals in reporting data to the interviewers or to the Fiscal Authorities. The number of interviewed taxpayers was 1098 (527 men and 571 women), as indicated in Table 1, but 66 taxpayers (6%) were not interviewed because they would not have been home for a long period of time during the survey period; for example, they may have been working outside the local area or abroad or the respondent may have simply stated this to avoid a subsequent interview or further contact. Therefore, they were eliminated from the analysis. There was one individual (1), a retiree, who had a missing value for SI and a zero value for FI. He/she was also eliminated from the analysis because the case was not interesting and because some variables involved in the models were not calculable. In fact, these latter variables were derived from a logarithmic function, in which the expression of its argument was zero. Among the remaining 1031 (=1098 –66 –1) units, as indicated in Table 5, 44 taxpayers (4%) declared zero income to the interviewers (SI=0) for different reasons: they misinterpreted the concept of income neglecting income from rental properties or land, they left out taxable benefits such as unemployment subsidies or more simply they made false statements. They were included in the data set and belonged to the categories of the unemployed (5), retirees (6), and other inactive persons (33).

Appendix B. *Independent variables and estimation details*

Each independent variable used in the multinomial logit model or in the regression model on Fiscal Income (FI) is defined below.

The *types of tax forms* utilized for reporting income were: the TF730, TF-Unico, and TF770. The TF730 form is more simplified and it is utilised by the majority of employed workers. The TF730 and TF-Unico contain details regarding sources of income (land, buildings, employment, and other taxable income) and relevant tax deductions and tax allowances (for a spouse, children, relatives, pension, etc.). The TF770 form, which is filed by employers for employees, contains information on the taxable income of employees and refers here to the potential FI of those who are exempt from filing a tax report (given that all the relevant information is provided by the employer). Therefore, this category should show lower FIs than the FIs of the other two categories. The TF730 was assumed as reference group (RG). Therefore, type of tax forms generated three binary or dichotomous variables: TF730, TF-Unico, and TF770. Each one assumed the value of 1 when the individual used the indicated type of form and 0 otherwise. The TF730 category was assumed as the Reference Group (RG).

Gender was transformed into the dichotomous variable "women", assuming 1 for women and 0 for men. The latter formed the RG.

Age and *Fiscal Income* (FI) were introduced into the model through a second-degree polynomial form (ax^2+bx+c) to capture some nonlinearities in the behaviours of individuals of different ages and FI values. However, the expected impact of age on earnings might have the same form, involving a high correlation between them: as young workers become older, their earnings will usually increase. Moreover, a significant effect of one of the two might incorporate the effect of the other. In the models, to have values of age and FI comparable with other regressors, which were binary variables, the original age values were divided by 50 and the original FI values were divided by 100,000. In other symbols, $x_{\mathrm{Age}} =$ Age/50 and $x_{\mathrm{FI}} =$ FI/100,000.

Education Level (EL) was summarised by the usual four categories, each one generating a dichotomous variable: primary (assumed as the RG), lower secondary, upper secondary, and tertiary education. Herein below, the generation of dichotomous variables will be implicitly assumed.

Activity Status (AS) was recoded into four categories: employed (assumed as the RG), unemployed, retiree, other inactive people. Note that the inactive category was split into two subgroups: retirees and other inactive people.

Employment Status (ES) was available divided into four categories: employed (assumed as the RG), entrepreneurship, self-employed, and business partner.

Skill Level on the job (SL) represented the occupational category as a combination of the professional position and skill level of the job. It was divided into in five levels: under-skilled (fixed as the RG), low-skilled, medium-skilled, high-skilled employees, and private or public managers.

Sector of Activity (SA) presented modalities based on the statistical classification of economic activities (NACE, Rev. 1.1 2002, from the French *Nomenclature statistique des Activités économiques dans la Communauté Européenne*, used until 2008), but with a slight modification/ adaptation for the 2001 Italian Census. They were grouped according to an ordinary categorisation. However, Section D (manufacturing) was split into two groups to obtain a suitable aggregation in keeping with the local economy. The first category was termed "1=Mechanical Engineering Industry" (MEI) and included the divisions (two-digit numerical codes) from 28 (manufacture of fabricated metal products, except machinery and equipment) to 35 (manufacture of other transport equipment). The second category was termed "2=Remaining manufacturing of Section D plus Sections C & E" (Remaining D+C+E) because it grouped the other divisions of Section D, Section C (mining and quarrying) and Section E (electricity, gas and water supply), given the local characteristics of these firms and their insignificant number. The third category referred to the modality "3=Trade & Transport" including Sections G, H, and I. The fourth category concerned "4=Services" and included Sections J, K, O, and P. The fifth category was expressed by acronym "5=PAEH" because it incorporated Public Administration (Section L), Education (Section M), and Health and social work (Section N). The sixth category was termed "6=Other sectors" (assumed as the RG) and contained various heterogeneous economic sectors such as agriculture, hunting and forestry, fishing, construction, and missing values. The reference group (RG) could have been constituted by missing values, given numerousness of the latter, but for the sake of uniformity with the standard procedure and models used in Sections 5 and 6 of this paper, the omission of one category was preferred even if it also included missing values.

Surveyed income was the main variable in the data analyses and corresponded to the total individual income, $y_{\mathrm{Tot};i}$, before tax allowances. The total taxable income was obtained from $y_{\mathrm{Tot};i}$, after deducting the allowances, the most important being the allowance for cadastral income from home ownership and the compulsory social security contributions for the self-employed. It presented peculiarities deriving from the planning of the ISCEMO1. The questionnaire used in the interview had many questions (variables) about income and wealth and, specifically, those concerning net earnings, real and personal estate, capital and financial assets over the last twelve months. The choice of this variable reference period was based on the knowledge that the longer away from the beginning of the time interval, the more confused or vague are recollections of income data (among others, Moore et al. 2000). Unfortunately, the interview period was from June 2002 to December 2002. Therefore, the beginning of the reference income period varied from June 2001 to December 2001. Such a variable reference income period among the subjects represented the most critical aspect for at least two reasons. First of all, hypothetical and nontrivial operations were necessary to obtain the individual yearly survey income (SI) to be compared with FI, but a detailed discussion of these operations and results is beyond the scope of the paper. Secondly, there is some evidence that the length of the recall period contributes to faulty recall (Moore et al. 2000). For example, it would have been better to refer to income over an annual period coinciding with the calendar year.

Some notes concerning the estimated model of FI are provided below.

R1. The analysis of residuals in the regression model, using the original FI as the regredend, showed heteroscedasticity and a non-normal distribution, which was right or positive skewed. The logarithm of FI weakened, but did not eliminate the heteroscedasticity and asymmetry, where the latter became significantly negatively skewed. Perhaps the dependent variable should have been transformed through a logistic function, which is rarely applied due to the difficulties in interpreting the results, or through a Box-Cox transformation with a particular lambda, also rarely applied due to difficulty in understanding the outcomes immediately. As stated in the body of the paper, all the available variables were included in the models. Except for age, they were qualitative variables and were transformed into binary or dummy variables having the value of 1 when they assumed the corresponding modality, and 0 otherwise, as specified above. Their regression coefficients expressed the variation of log(FI) with respect to the constant denoting the reference group (individual/ category): taxpayer using the TF730 form, men, at the age of zero, with Italian primary school education level, employed, under-skilled, working full time, with a contract of unlimited duration, without paid employees (which is the value equal to zero of the dummy variable "with paid employees"), and in a residual group of economic activities (and in the sample of unused reserves for the model using the full sample).

R2. Two variables, upper secondary education level and self-employed, showed significant increases in FI in the sample of interviewees with respect to the full sample. Upper secondary education level contributed with 13.4% given by the difference between the corresponding coefficient of the interviewees (0.508) and the coefficient of the full sample (0.374), i.e., 0.508−0.374= 0.134, which in percentage terms became 13.4% (Table 4). The impact of self-employed was obtained in the same manner: −0.338 – (−0.553)= 0.215, or, expressed as a percentage, 21.5%.

R3. The vertex of the quadratic function $(ax^2 + bx + c)$ is given by $x_T = -b/2a$, representing the turning point. Therefore, in Table 4, the observed age coefficients revealed a substantial increasing trend. For example, in the sample of interviewees, the increase of FI with age turned around at $x_T = -2.116/[2 \times (-0.573)] \approx 1.85$ corresponding to Age= $x_T \times 50$= 1.85×50= 92.5 years. Note that multiplication by 50 is needed to obtain the value in years of age, as the variable introduced in the model was divided by 50. Perhaps the decreasing effects were captured by other variables, such as activity status and employment status, even if an easy, but uncertain interpretation might indicate a gradual impoverishment of the younger generations and the tendency of the older generations to preserve their rights and privileges.

***Appendix C**. Details on over- and under-reporting*

There were many reasons for assuming FI as the reference income. For example, if the taxpayer is not a tax evader, this income is verified and more precise than the amount of income collected through a survey. Employees in the public sector cannot be considered as evaders, constituting a useful subsample to evaluate measurement errors. In this sense, FI was the reference and correct income. Obviously, although it is unusual, it is possible to assume SI as reference income too.

Looking at Table 5, one may note some unexpected or surprising data. Further clarification is provided on the example mentioned in the text: there were 76 self-employed men, showing means of $\overline{\mathrm{SI}}$ = € 31,287 and $\overline{\mathrm{FI}}$ = € 42,939, but $\overline{\%\,\mathrm{D}}$ = 78.4%. At first sight, it is not clear why the mean of the individual percentage differences, $\overline{\%\mathrm{D}}$, among the men was so high. The minimum of $\%\mathrm{D}_i$ was −96.5% (but the minimum cannot be lower than −100%) and the maximum of $\%\mathrm{D}_i$ was 4065.0%. Among 76 self-employed men, there were 20 over-reporters with a minimum of 2.7% and a maximum of 4065.0%. Only 9 over-reporters had $\%\mathrm{D}_i$ values greater than 100%, with a minimum of 119.6% and a maximum of 4065.0%. To provide an example, the data relative to the maximum of $\%\mathrm{D}_i$ for one of these 9 individuals are listed here: SI' = € 20,325.13 and FI' = € 488 and thus $\%\mathrm{D}'$ = 4065.0%, given by $100 \times (20{,}325.13 - 488)/488$. This difference does not necessarily indicate evasion, but it might be a result of costs reducing taxable income, tax deductible costs, and other particular favourable cases provided for by law. Therefore, these observed values were possible outcomes, but resulted in highly influential cases, which posed a quandary: eliminate them from the dataset, equalize them in some way or leave them as they are. The latter option was chosen. In conclusion, there were fewer over-reporters than under-reporters, but their $\%\mathrm{D}_i$ values were so high that they led to a positive mean, $\overline{\%\mathrm{D}}$, even if the category showed a lower mean of SI with respect to the mean of FI.

The mean of percentage differences for all interviewees (total in Table 5) was high and positive, although not statistically different from zero. However, as explained above, this may be misleading because the percentage difference had a lower bound equal to −100% and no upper bound. In the sample of all interviewees, 39 individuals showed a percentage difference often exaggeratedly greater than 100%, ranging from 100.3% to 33,876.6%. On the one hand, women who were entrepreneurs, self-employed, labourers or retirees tended to be over-reporters. On the other hand, women who were public officials or executives, employees, unemployed, or inactive tended to be under-reporters. The behaviour of men was slightly different across the employment status and job title categories. An interesting profile emerged for the categories less prone to evasion: official executives (79), employees (258), labourers (128), and retirees (341). They constituted a subsample of 806 units, named “restricted sample”, which was smaller than the sample of interviewees and suitable for an analysis of measurement errors. Men and women public officials or executives had the same tendency to under-report income, but women showed a higher percentage difference with respect to men. Among employees, labourers, and retirees an opposite tendency emerged between women and men. In certain categories men were over-reporters and women were under-reporters, while in others, an opposite behaviour was observed, leading to a compensation of the under- and over-reporting.

Two details (D) on the distribution of errors are reported below.

D1. The distribution of the logarithm of errors is shown in Figure 3. For over-reporters (308 units), the mean (μ) was 7.37 and the standard deviation (SD) was 1.54, appearing to be mildly leptokurtic (4.60) and negatively skewed (–0.75). Similarly, for under-reporters (498), the mean was 7.86 and the SD was 1.31, appearing to be mildly mesokurtic (3.40) and negatively skewed (–0.32). It can be noted that the number of under-reporters (61.8%) was higher than the number of over-reporters, confirming that false statements concerning income are more frequent than true statements. Moreover, as expected, the mean of under-reporters (μ=7.86) was higher and the distribution was more concentrated (SD=1.31) than those for over-reporters (μ=7.37, SD=1.54, respectively), revealing a tendency among respondents to conceal their income in an interview.

D2. The distribution of the logarithm of percentage difference for surveyed and fiscal income is shown in Figure 4 and had the same pattern. For over-reporters, the mean (μ) was 2.31 and the standard deviation (SD) was 1.62, appearing to be mildly leptokurtic (4.81) and slightly negatively skewed (–0.10). For under-reporters, the mean was 2.59 and the standard deviation was 1.11, appearing to be mildly mesokurtic (3.31) and negatively skewed (–0.57). In the distribution, there was also a truncation on the right deriving from the impossibility of under-reporting to attain a value lower than −100%, while it was possible to over-report a value greater than 100%. Again, as expected, the mean of under-reporters (μ=2.59) was higher and the distribution more concentrated (SD=1.11) compared to those of over-reporters (μ=2.31 and SD=1.62, respectively). The distribution of under-reporters was bimodal and the lower mode, specifically the right spike, depended on the threshold of −100% which cannot be exceeded.

***The variables* used as regressors** in the regression models for interviewees were collected during the survey, while the regressors of the previous models (in Table 3 and Table 4) came from the Census of 2001. However, some regressors in Table 6 were similar to those of the Census and had the same definition illustrated in Appendix B: type of tax form, FI, gender, age, education level, sector of activity. Age and FI were introduced into the model in a second-degree polynomial form again for the same reasons.

Employment Status (ES) was combined with job title or skill level of the job and the Reference Group (RG) consisted of labourers. This combination was easily generated with the survey data, while it was not possible to obtain this variable with the census data.

Tenure Status of Households (TSH) distinguished between owners (RG), tenants or subtenants paying rent at prevailing or market rate, free accommodations, and other special situations included in the RG.

Degree of Relationship between Family members (DFR) referred to: head of a family (RG), partners, daughters or sons, and other types of relationships included in the RG.

Marital Status (MS) included four categories: single, married (RG), divorced, and widowed.

Non-Cooperation for SI was a binary variable indicating people who refused to give information about SI; in that case its value was 1 and 0 otherwise. This binary variable captured only the few cases (6 individuals in the restricted sample of 806 units), who provided false statements, stating that their SI was equal to zero: errors equal to −100%.