INTERNATIONAL JOURNAL OF MICROSIMULATION (2013) 6(1) 86-121
INTERNATIONAL MICROSIMULATION ASSOCIATION

The importance of choosing the data set for tax-benefit analysis

Lidia Ceriani

PAM, Bocconi University and Econpubblica.
Via Rontgen 1, 20136 Milano (IT);
e-mail: lidia.ceriani@unibocconi.it

Carlo V. Fiorio

DEMM, University of Milan and Econpubblica.
Via Conservatorio, 7, 20122 Milano (IT);
e-mail: carlo.fiorio@unimi.it

Chiara Gigliarano

DISES, Università Politecnica delle Marche, Ancona.
Piazzale Martelli 8, 60121 Ancona (IT);
e-mail: c.gigliarano@univpm.it

ABSTRACT: Given the increased availability of survey income data, in this paper we analyse the pros and cons of alternative data sets for static tax-benefit microsimulation in Italy. We focus on all possible alternatives, namely using (a) SHIW or (b) IT-SILC data using a consistent net-to-gross microsimulation model, or (c) IT-SILC data using the gross incomes provided since 2007. Our results suggest that IT-SILC improves in the regional representativeness of the Italian population and does not perform worse than SHIW as for most demographic characteristics, SHIW provides more information regarding building and real estate incomes. Gross income variables simulated by using the net-to-gross module included in the TABEITA microsimulation model and calibrating for tax evasion provide a very precise fit with external statistics. Simulated IT-SILC gross income data fit external aggregate data even better than gross income data provided in IT-SILC, which tend to largely overestimate self-employment income. Finally, we suggest to match IT-SILC with SHIW to include in the former the information on building and real estate incomes that are contained only in the latter. This allows us to reach a very satisfactory validation of the final data set.

KEYWORDS: H20, H24, H26, C63.

JEL classification: Tax-benefit microsimulation, Italy, TABEITA, SILC, SHIW.

1. INTRODUCTION

Many years have elapsed since Atkinson and Sutherland (1983) advocated the use of survey based data set for the development of microsimulation models. They analysed the family composition and circumstances recorded in the 1980 UK Family Expenditure Survey with the hypothetical family types used in the Department of Health and Social Security (DHSS) tax-benefit model, finding that only 4% of actual family types were covered by the assumption of the complex DHSS hypothetical family model. Although models based on hypothetical families are still developed and maintained (e.g. OECD, 2010), since the late 1980s virtually all developed countries have at least one model to simulate changes of taxes and benefit on individual incomes (see among others Mitton et al., 2000; Gupta and Kapur, 2000; Sutherland, 1995) and after the transition from centralized planning to a market economy and the consequent need of financing government policy through taxation, also Eastern and Central European countries showed an increased interest in tax-benefit microsimulation models (see for instance Coulter et al., 1998; Juhàsz, 1998; Lelkes, 2007; Lelkes and Sutherland, 2009). In recent years European policy makers have also developed strong interest in the development of harmonized microsimulation models, funding the EUROMODupdate (EMup) project for updating and extending EUROMOD. EUROMOD is a EU-wide microsimulation model aimed to provide estimates of the distributional impact of changes to personal tax and transfer policies taking place at either the national or the European level, to assess the consequences of consolidated social policies, and to understand how different policies in different countries may contribute to common objectives (Sutherland and Figari, 2013).

However, the quality of a microsimulation model is strongly related to the availability of good quality data (Sutherland, 1991). At present, although very few national microsimulation models can be developed on administrative data, which are characterized by very little measurement error as opposed to survey data, microsimulation modellers are often in the happy position of being able to choose among different survey data sets, which differ for the way data are collected, and the type of information that is collected. This is in fact what happens in Italy.

Surveys of reasonable size collecting information on income data and other individual and household characteristics for a representative sample of Italian households have been made available by the Bank of Italy for over 25 years with the Survey of Household Income and Wealth (SHIW) and, since 2004, by the Italian Institute of Statistics (Istat) with the Italian version of the Statistics on Income and Living Conditions (IT-SILC). Both SHIW and IT-SILC are suitable data sets for microsimulation modelling. However, as they provide income data net of all taxes and

social contributions paid, a net-to-gross procedure is required to simulate before-tax income prior to any tax-benefit simulation is performed. Since 2007, Istat complemented the IT-SILC data with gross income data partly simulated and partly produced by exact matching of survey data with administrative ones (see Istat, 2009; Istat, 2011).

Italian microsimulation modellers, therefore, can currently choose among three different sources of gross incomes for analyzing distributional effects of tax-benefit models. They can use a microsimulation model which produces gross incomes from the net-to-gross conversion of disposable incomes recorded in either SHIW or IT-SILC data, or they can start from gross income data provided in IT-SILC.

This paper aims at clarifying pros and cons of these alternative data sets for microsimulation analysis in Italy[1], but it is of general interest also for the development of microsimulation models in other countries. We first suggest that where more than one survey exists, one should be aware of the pros and cons over different dimensions of available surveys to make use of the best one depending on the research question addressed and that an assessment of standard errors of the main statistics considered could also be an important element to consider. By noticing that gross incomes recorded in SILC data (which are often used as input data in microsimulation models across Europe and for the EUROMOD project) provide information on taxes paid on aggregate income variables, whose components are sometimes treated differently by the tax code, we suggest that having a consistent-over-time net-to-gross procedure would allow both more detailed analyses of tax-benefit policies and the use of input data from complementary data sources even in case gross income values were not provided. We also describe the correction of tax evasion of self-employment income as part of the net-to-gross procedure. This is an improvement with respect to a rough correction of the gross income values, which is often used when a net-to-gross procedure is not implemented. Finally, we also suggest that one could make use of matching techniques to match different surveys and enrich the input data used for microsimulation models. This is in fact what we finally suggest here to enrich IT-SILC data with information regarding the value of second homes and other real estate and buildings coming from the SHIW data.

The structure of the paper is as follows. First of all, we will describe the available Italian datasets focussing on 2008 incomes to avoid updating procedures which could introduce estimation biases. We will discuss their sample design, unit and item non-response, post-stratification weights and other quality issues (Section 2). Then we will briefly present TABEITA, the static microsimulation model used here, focussing on the net-to-gross module. We have been

developing TABEITA for over a decade and it is currently used to provide gross income data for the EMup project mentioned above (Section 3). Hence, we undertake an assessment of both datasets focussing on socio-demographic variables and comparing them with statistics produced by external sources when available and on net income and their distribution in the Italian population (Section 4). Finally we also compare simulated gross income data with gross income variables available in IT-SILC and generated by Istat by using microsimulation and exact matching with administrative data. For the main statistics computed we also provide their standard errors. We show that simulated gross income data using IT-SILC net income data show the best fit with aggregate external statistics, especially when integrated with information on income for other building and real estate earnings (Section 5). Finally some conclusions are drawn (Section 6).

2. SHIW VS. SILC: A GENERAL DESCRIPTION

At present, researchers aiming to analyse disposable income distribution in Italy can choose among two alternative datasets: the Survey on Household Income and Wealth (SHIW) and the European Statistics on Income and Living Conditions (EU-SILC).

SHIW is provided by the Bank of Italy and was available with yearly frequency from late 1960s to 1984 and about every two years since then (see Bank of Italy, 2010 and Brandolini, 2008). Since 1986, it collects information about roughly 8,000 households and 20,000 individuals. This database has been widely used to study the economic behaviour of Italian households and is included in the Luxembourg Income Study (LIS), a cross-country project aimed at producing a harmonized social and economic data from household surveys independently collected in about 45 countries.

EU-SILC has been conducted yearly, starting from the year 2004, in 25 EU member States, plus Norway and Iceland, with the coordination of Eurostat (for an analysis of strengths and problems of EU-SILC, see Iacovou et al., 2012 and Eurostat, 2007). It focusses on each country's resident households and has the purpose of collecting both quantitative information concerning income received by all household members and qualitative indicators on living conditions of households and individual quality of life (such as educational level, health status, house quality and occupational status, conditions of financial stress).

In this paper we select only the waves of SHIW and of the Italian version of EU-SILC (henceforth, IT-SILC) recording incomes earned in 2008. The former collects information about

7,977 households, corresponding to 19,907 individuals and 13,268 income earners. Households have been sampled from the registry office list of municipalities and households were interviewed between January and September 2009. IT-SILC contains all standard EU-SILC variables and a long list of additional ones, which are relevant for the development of microsimulation models. The survey used in this paper (2009 IT-SILC, recording incomes earned in 2008) collects information about 20,492 households and 51,196 individuals interviewed between September and October 2009. Both SHIW and IT-SILC are developed to be representative of the national population of households. Although Eurostat directives on EU-SILC require only national representativeness, IT-SILC has been designed to ensure representativeness at the regional level.

A peculiarity of IT-SILC is that information on disposable income is integrated with data coming from administrative sources, such as Istat National Accounts, the Department of Finance of the Ministry of the Economy and Finance, the National Institute for Social Security Register, and from the Labour Force Survey, with the purpose of reducing measurement errors, increasing coverage and data accuracy, minimizing the under-reporting of incomes (especially self-employed income) and optimizing the imputation process of missing values. This operation is based on exact matching between individual sampling data and administrative records, using the tax code as matching key.[2] Starting from the year 2007, a relevant improvement was introduced in IT-SILC, namely gross incomes are provided along with net ones, consistently with all other countries surveyed in the EU-SILC. Gross incomes are not collected during interviews but recovered by exact statistical matching with tax records using the individual fiscal code as matching key or by using the Siena Microsimulation Model for variables which could not be recovered otherwise (for more details, see Betti et al., 2011; Istat, 2011).

2.1. Sample design, non-response and other quality issues

The sample design of both datasets is based on a two-stage sampling, according to which a sample of municipalities is selected, followed by a sample of households from each selected municipality. Before selecting the first stage units, municipalities are stratified according to region and demographic size. Within each stratum, the sampling scheme includes all the municipalities above a certain population, while the remaining municipalities are selected so that larger municipalities have higher probability to be included in the sample. In the second stage, for each of the selected municipalities, households are randomly chosen from the registry list.

When comparing the quality of different statistical surveys a relevant criterion that should be taken into account is accuracy, usually characterized in terms of error in statistical estimates (bias

and random error), or described in terms of the major sources of error that potentially cause inaccuracy (e.g., coverage, non-response; see, among others, Statistics Canada, 2003). Unit non-responses occur when no interview can be obtained, while item non-responses when only some of the questions are not answered. Both IT-SILC and SHIW data suffer from these two types of non-response, although in different degrees.

The unit response rate in SHIW is 56.1%, mostly due to households' unavailability. According to the Bank of Italy (2010), the difficulty of getting an interview increases with income, wealth, education of the household head and household size, and it is lower for families living in the South or in the Islands, in small municipalities, and those whose head is older or unemployed. The amount of item non-response occurring in SHIW is negligible, as shown in Bank of Italy (2010). In particular, most of the missing values relate to non-monetary incomes for employees and monetary income for the self-employed. For these variables, item non-response corresponds on average to less than 4 per cent of the sample (for a more detailed discussion about non-response behavior in SHIW see D'Alessio and Faiella, 2002). The methodology employed for the imputation of missing values in SHIW is based on regression models, adding a random component drawn from a Gaussian-distributed population with zero mean and variance equal to the variance of the regression residuals to avoid over-concentration around the mean.

As far as IT-SILC is concerned, the unit response rate is 85.50% and the item non-response rate is not negligible. It is around 11.1% on average, but it varies across different items, from 3.5% for net employment income to 9.8% for net gains from self-employment, to 11.6% for pension from work, and 23.3% for financial income (for details, see Table 4.4 in Istat, 2008a).

Interestingly, the Bank of Italy's interview strategy excludes a priori those families who refuse to provide information about their incomes. In this way, the Bank of Italy puts higher priority on minimizing the item non-response for income variables, at the expenses of a smaller participation rate. Istat, on the contrary, pursues the aim of maximizing the participation rate, so allowing for higher probability of item non-response, to be partly remedied by exact matching with administrative sources (see Istat, 2008a).[3]

In general, one cannot claim that higher response rates imply smaller non-response error. The literature provides many examples of this (see, among others, Groves and Peytcheva, 2008, Schouten et al., 2009 and the references therein). In particular, Andrigde and Little (2011) find that the impact of non-response on bias depends primarily on three factors: the non-response rate, the strength of the proxy variable in predicting a variable of interest, and the difference in

proxy means for respondents and non-respondents. One of their main findings is that, if non-response is highly correlated with the variable of interest, there can be a large bias even with a high response rate. A deep analysis of the quality of the two datasets is beyond the scope of this paper. For some recent methods for analyzing survey non-response bias, we refer, among others, to Wagner (2010) and Andridge and Little (2011) and the reference therein.

Differential response rates are taken into account in the definition of sampling weights and ultimately of post-stratification weights, which are relevant for obtaining aggregate statistics and for policy analysis. In SHIW, post-stratification weights are obtained as follows: first a sampling weight is computed as the inverse of the probability of inclusion of the household in the sample. This weight is then corrected for the overall non-response rate, multiplying it by the inverse of the response rate of the municipality where the unit lives. Finally, this weight is calibrated using statistics about the population on a set of variables (namely, gender, age groups, three main geographical areas – North, Center, South and Island – and size of the municipality of residence), and all member of the same household are given the same weight.

The IT-SILC weights are developed in four steps (see Istat, 2010b for details). The initial weights are the inverse of the inclusion probability of each household. The second step contains a slight correction for non-response both at the household and the individual level. Hence, correction for over/under representation of certain population groups is implemented through calibration by age, gender, size of resident population at the NUTS II-level.[4] Finally, a calibration is performed to make sure that all members in the same household receive the same weight.

Data quality is also strongly related with the process of data production. The literature has shown that interviewers play a crucial role in determining the data quality. The level of experience of the interviewers and the fieldwork quality controls on interviewers are all relevant issues. However, we are not able to collect such information since it is not publicly available. We can only observe that in SHIW data collection was mainly through the CAPI (Computer-Assisted Personal Interviewing) procedure, which allows for prompt controls of possible inconsistent answers. The remaining interviews are made through papery questionnaire (PAPI, Paper-And-pencil Personal Interviewing), run by professional and high-educated interviewers (Bank of Italy, 2010). In IT-SILC data collection has been done using only the PAPI method with periodically trained interviewers (Istat, 2008a). However, we are not in a position to assess which method is actually delivering better quality data.

Table 1 provides a summary of the main differences between SHIW and IT-SILC.

Table 1 Comparison between 2008 SHIW and 2009 IT-SILC

	SHIW	IT-SILC
Provider	Bank of Italy	Istat
Frequency of data collection	Biannual	Annual
Period of data collection	January-September 2009	September-October 2009
Income reference period	2008	2008
Coverage	Private households	Private households
Statistical units	Households and individuals	Households and individuals
Sample size (households)	7,977	20,492
Sample size (individuals)	19,907	51,196
Response rate	56.1%	85.5%
Geographical Representativeness	Three Macro regions	Regional
International comparability	No[1]	Yes
Sample design	Two-stage (first stratified)	Two-stage (first stratified)
Data collection methodology	CAPI/PAPI	PAPI
Gross income variables	No	yes (starting from 2007)

Source: Our calculation and analysis of SHIW and IT-SILC.
Note: [1]International comparability is not among the main aims of the SHIW. It can however be obtained upon adequate harmonization of main variables, such as the "lissification" procedure (see LIS project).

3. THE TABEITA MICROSIMULATION MODEL

The Microsimulation Model used in this paper is TABEITA (Tax Benefit model for ITAlian personal income taxation) developed using the 2008 tax rules. This model belongs to the family of static tax-benefit microsimulation models which have been developed and regularly maintained since 2001 at Econpubblica at Bocconi University (Fiorio, 2009). Up to 2006 TABEITA models have been built and maintained using SHIW data, the only available source for tax-benefit analysis. Since 2006 TABEITA has been developed using IT-SILC as the main dataset for two main reasons. Firstly, our comparative analysis of SHIW and IT-SILC data suggested us to prefer the latter, as discussed as discussed in this paper. Secondly, because the EUROMOD model, which uses gross data simulated using TABEITA, requires us to use SILC data for harmonisation of results across the EU. However, for this project we harmonized SHIW and IT-SILC so that either data set can be used for the very same version of TABEITA model.

As survey data for tax-benefit analysis are collected in interviews where respondents are asked only about their disposable income, the first aim of microsimulation models such as TABEITA is to simulate the Italian Personal Income Taxation (PIT, known in Italy as Irpef) and social security contributions. TABEITA can be described as a deterministic transformation of a given sample into a new one. Let $\boldsymbol{y}^A$ and $\boldsymbol{y}^B$ be the vectors of after-tax (AT) and before-tax (BT) income, respectively: the former vector is obtained from the latter through a tax function, say τ_i ,

i = 1, 2, …, N, where N is the number of individuals in the sample. Since the data are recorded net of taxes and social contributions, the first role of the model is to provide a net-to-gross procedure to simulate individual BT income:

$$y_i^B = \tau_i^{-1}(y_i^A) \tag{1}$$

for all i=1,...,N. There are two major complications here. First, the tax transformation τ_i is not the same for all individuals. Personal income taxation in Italy is on an individual base; the amount of the tax one has to pay depends on the type of incomes she receives and her family characteristics. For instance, arrears do not enter the PIT base: they are taxed with a proportional tax rate while work and pension income is taxed with progressive tax rates; tax credits depend on a set of individual and family characteristics, such as the number of dependent children, whether the spouse is dependent, whether income comes from self-employment, employment or pension, etc. Secondly, the tax transformation in (1) is only piecewise linear and its inversion requires the exact identification of the interval where y_i^A falls in. The tax transformation, τ_i, used to recover y^B is obtained from the tax code assuming full take-up rates of tax credits and tax deductions.[5]

TABEITA is developed in six consecutive steps. The first module involves the preparation of the dataset and data consistency checks. This is the only step which is specific to the input data used (e.g. either SHIW or IT-SILC) and its aim is to produce an input data set with all information required by TABEITA in the following five steps. Whenever an income variable, relevant for tax-benefit microsimulation, is contained in a dataset but not in the other it is set equal to zero in the latter or calibrated clarifying the assumptions introduced. The second step involves the generation of tax units. Tax units are generated identifying the household head[6], linking him/her with his/her spouse (if present and regardless of his/her income) and dependent relatives using information on family relations. Independent relatives[7] belonging to the same household are defined as different tax units, and their family relationships reconstructed as above exploiting the rich information available in the input data sets. The third step involves the identification of different components of individual disposable income and their relative share.[8] The fourth step randomly assigns tax deductions and tax credits to eligible individual taxpayers where no information is available[9] based on aggregate statistics provided by the Department of Finance, Ministry of Economics and Finance (henceforth MEF). The fifth performs the net-to-gross tax function inversion as in (1). Finally, the output validation analysis is performed using external statistical sources.

As the structure of Irpef is standard[10] let us describe TABEITA in general terms.

Let Y_j be the tax debt at the lower bound of the tax bracket *j*, with *j*=1,2,..., *J*, T_j be the lower bound threshold of tax bracket *j*, t_j be the tax rate, y^B be the income before tax and net of exempt income, and *tc* and *ta* be tax credits and tax allowances, respectively, one can write the after tax income (y^A) as:

$$y^A = y^B + tc - Y_j - t_j \cdot (y^B - ta - T_j) \tag{2}$$

which implies that

$$y^B = \frac{y^A + Y_j - t_j \cdot (ta + T_j) - tc}{1 - t_j} \tag{3}$$

Adding the additional regional personal income tax, which is defined on the same tax base as Irpef and is different depending on the region a taxpayer resides, and denoting by Q_j the tax debt at the lower bound of tax bracket *i*, by Z_j the lower bound threshold of tax bracket *j* and by r_j the tax rate of tax bracket *j*, one can obtain the before-tax income as:

$$y^B = \frac{y^A + Y_j + Q_j - t_j \cdot (ta + T_j) - r_j \cdot (ta + Z_j) - tc}{1 - t_j - r_j} \tag{4}$$

This step involves the endogenous simulation of major tax deductions (e.g. social security contributions for self-employed) and tax credits (e.g. for family burdens, and type of income earned, if by employment, self-employment or pension) through a number of iterations up to the point where an exit condition is reached.[11]

Finally, the last step simulates BT income by income source by adding to each AT income source a share of the total tax based on the estimation of income shares as defined in step three above. Adding also exempt incomes one obtains total disposable income.

3.1. Calibration of tax evasion

It is a matter of fact that regardless the income data source used for TABEITA (as well as for all microsimulation models developed on Italian data that we are aware of) and of the year considered, simulated aggregate self-employment income is consistently larger than the aggregate amount recorded by administrative sources. Various interpretation of this could be suggested. Consistently with the tradition of microsimulation with Italian data (see among others, Mantovani, 1998; Bernasconi and Marenzi, 1999; Fiorio and D'Amuri, 2005), we assume here that the difference comes from the fact that self-employed workers declare in an anonymous interview a higher level of income as opposed to what they actually declare to tax authorities because they have no incentive to hide it. This is a debatable assumption but it simply comes

from observing a regularity in the Italian data. Moreover, this assumption does not necessarily contrast with findings of other authors suggesting that self-employed workers declare in anonymous surveys a lower level of disposable income that employees with similar consumption patterns (Pissarides and Weber, 1989), which we do not address here and would simply increase the extent of income concealment by self-employed. Hence, using available external statistics we calibrate reported income of self-employed taxpayers defining a uniform rate of unreported self-employment income. The amount of this unreported income is then subtracted from the tax base (3), as if it was a tax exempt component of disposable income, and the simulated tax is computed as described above. Disposable income is finally obtained adding to AT income all exempt incomes including calibrated unreported income from self-employment, which we call "tax evasion". Depending on the year and the survey data used, the calibrated coefficients in TABEITA range between 20% and 30%, which is also remarkably close to estimates of tax evasion for self-employment income provided by Istat (2010a). This approach could be further improved for instance estimating differential tax evasion rates by income groups or regions, but we preferred to keep it simple in the current application.

4. SHIW VS. SILC: INFORMATIONAL DIFFERENCES

We now move on to analyse the differences between SHIW and IT-SILC in terms of the information provided, looking first at variables that are contained in a dataset but not in the other, then at their ability to represent socio-demographic characteristics of the Italian population and finally at the analysis of net income data.

4.1. Informational differences

Although the two data sets are characterized by a similar structure of the questionnaire and a number of similar questions of relevance for tax-benefit microsimulation models, some variables are contained in SILC but not in SHIW, and vice versa. In particular, IT-SILC only provides detailed information on (*i*) family, maternity and education-related allowances, unemployment benefits, social assistance at the municipal level, benefits for mortgage and for rent; (*ii*) house-related expenses, such as condominium fees, heating, electricity, tap water, garbage, telephone, gardening and ordinary maintenance house expenses, as well as the local property tax and (*iii*) specific sources of income, such as employment income arrears, fringe benefits, Christmas bonus, additional payments based on productivity, supplementary pensions, employment additional compensations, income from rent of rooms or accessory structures (such as cellars or garages) of the main residence, interests and dividends from capital investment. SHIW does not

provide such detailed information in distinct variables and often this information is aggregated to other variables (e.g, family allowances and unemployment benefits are provided together with employment income[12] For a detailed description of the additional variables available in IT-SILC, see Ceriani et al. (2011, 2012).

On the contrary, SHIW data include information about (*i*) education expenses, such as childcare, primary school and university fees; (*ii*) daycare and elderly care expenditures; (*iii*) health-related expenses, such as diagnostic exams and private health insurance; (*iv*) donations (to ONG and liberal); (*v*) refurbishing expenditure allowances; (*vi*) life insurance; (*vii*) pension arrears and private pension plans. Moreover, SHIW data provide a detailed set of information about (*viii*) means of savings and (*ix*) building and real estate properties different from the main residence.[13] Note that whereas in SHIW a lot of information about households' financial wealth is provided, no information is given about the income flows that it generates. On the other hand, IT-SILC only records the amount of income received from interests and dividends from capital investment. However, neither of them is used in the simulation of the PIT because they are subject to separate taxation and do not enter the PIT base.

On the contrary, the informational advantage of SHIW vs. IT-SILC is large as for non-rented other properties. In fact, both datasets provide general information about homes and rented properties, in particular regarding the rent that households think they should have paid, have they not owned their homes (imputed rent), which is used to calibrate cadastral income that enters the tax base for 2008 PIT. However, while IT-SILC provides only a yes/no variable regarding the fact that a household owns other properties without renting them out, SHIW provides also detailed information about how much could be the market value of these properties, which again is used to calibrate the cadastral value entering the PIT base.

In fact, the stock of second homes and other properties is large in Italy. Table 2 shows that out of all these properties, 52% is made of residential building units, 26% of arable land, 7% of garages and basements. Out of a total of nearly 4,000 units of second homes and other buildings declared in SHIW, less than one third is rented providing a taxable income. The rest enters the tax base as cadastral income only and is mostly made of second homes. Notwithstanding the concerns about the precision of imputed rent for predicting cadastral values of real estates and other buildings, this is the only information available for Italian tax-benefit microsimulation and ignoring it would produce a larger bias in the estimated gross income.

Table 2 SHIW additional Information: Real-Estate Properties different from principal residence - Average yearly effective and imputed rent

	Frequencies %	Rented properties Effective Rent (c) Euro	Unrented properties Imputed Rent (d) Euro	(d/c)-1*100 Diff. %
Total units of building and real estate	100	5,217	2,610	-49.97
Residential buildings	51.53	4,985	3,528	-29.22
Offices	1.46	7,914	5,686	-28.15
Storehouses	4.21	7,665	4,734	-38.23
Shops	4.11	7,195	5,743	-20.18
Workshops	0.52	4,860	11,892	144.68
Garages and basements	7.45	3,179	1,280	-59.74
Arable land	25.56	2,570	1,146	-55.39
Non-arable land	5.15	4,350	856	-80.33

Source: our elaboration on SHIW data.

4.2. Demographic and socio-economic non-income variables

With respect to total population, both SHIW and IT-SILC weighted samples sum up to the Italian total population and by gender with little error (see Table3, Panel a).[14]

Table 3 Population Statistics: frequencies and their standard errors

	External source Frequencies (a)	SHIW Frequencies (b)	SHIW Standard error	Diff. % (b/a-1)*100	IT-SILC Frequencies (c)	IT-SILC Standard error	Diff. % (c/a-1)*100
Panel a: total population							
Total	60,045,068	59,721,994	425,204	-0.54	59,726,481	187,436	-0.53
Males	29,152,423	29,039,856	298,942	-0.39	29,034,258	133,299	-0.41
Females	30,892,645	30,682,138	302,382	-0.68	30,692,223	131,737	-0.65
Panel b: Population by Geographical Partition							
North West	15,917,376	15,832,738	243,504	-0.53	15,821,674	99,431	-0.60
North East	11,473,120	11,351,612	162,575	-1.06	11,387,593	63,830	-0.75
Center	11,798,328	11,708,006	192,699	-0.77	11,720,956	76,598	-0.66
South	20,856,244	20,829,638	237,817	-0.13	20,796,258	116,865	-0.29
Panel c: Population by Regions							
Piemonte	4,432,571	4,040,054	69,762	-8.86	4,404,240	43,130	-0.64
Valle D'Aosta	127,065	213,437	9,954	67.97	128,551	1,773	1.17
Lombardia	9,742,676	9,877,494	209,934	1.38	9,684,939	64,349	-0.59
Trentino Alto Adige	1,018,657	926,708	42,773	-9.03	1,018,870	13,472	0.02
Veneto	4,885,548	4,985,755	122,846	2.05	4,849,150	34,935	-0.75
Friuli Venezia Giulia	1,230,936	1,268,878	30,745	3.08	1,221,728	15,022	-0.75
Liguria	1,615,064	1,701,753	44,885	5.37	1,603,944	26,705	-0.69
Emilia Romagna	4,337,979	4,170,271	87,184	-3.87	4,297,845	33,274	-0.93
Toscana	3,707,818	4,064,424	81,482	9.62	3,678,686	34,202	-0.79
Umbria	894,222	899,923	26,000	0.64	893,152	13,906	-0.12
Marche	1,569,578	1,632,901	35,183	4.03	1,559,290	18,397	-0.66
Lazio	5,626,710	5,110,758	146,875	-9.17	5,589,828	46,453	-0.66
Abruzzo	1,334,675	989,980	37,424	-25.83	1,322,333	22,113	-0.92
Molise	320,795	563,074	27,802	75.52	319,701	5,599	-0.34
Campania	5,812,962	5,121,838	125,890	-11.89	5,796,817	52,115	-0.28
Puglia	4,079,702	3,859,146	106,992	-5.41	4,074,872	41,321	-0.12
Basilicata	590,601	1,621,012	59,568	174.47	590,344	10,574	-0.04
Calabria	2,008,709	1,990,086	51,394	-0.93	2,008,038	30,027	-0.03
Sicilia	5,037,799	4,811,870	126,081	-4.48	5,024,861	62,789	-0.26
Sardegna	1,671,001	1,872,632	36,650	12.07	1,659,292	35,834	-0.70

Source: our elaboration on DemoIstat, values as 1st January 2009 (External Source) and SHIW and IT-SILC data.
Note: Following EUROMOD conventions, in IT-SILC individuals who were born after the income reference period have been dropped (see Ceriani et al., 2012).

Similarly, as for the geographical partition (Table 3, Panel b). This comes at no much surprise as an outcome of the way both datasets calibrate grossing-up weights.

At the regional level IT-SILC performs better than SHIW (see Table 3, Panel c). This is due to the fact that SHIW is representative of the resident population by macro-regional partitions, whereas IT-SILC by regions. While IT-SILC discrepancies with respect to external sources range from -0.9% (Abruzzo and Emilia Romagna) to 1.2% (Valle d'Aosta), SHIW range from -25.8% (Abruzzo) to 174.5% (Basilicata). Although the SHIW data set guidelines clearly discourage the use of SHIW data for regional analysis, it has often been used as the only available data set: these descriptives statistics substantiate concerns regarding the use of SHIW for regional analysis.

Moving on to the analysis of the population sample by age groups, we find that SHIW dataset overestimates by roughly 10% the older age group and it underestimates by roughly 6% the younger group (Table 4). As for the age distributions within different geographical partitions, we find considerable discrepancies between survey data and population totals using the SHIW sample. On the other hand, IT-SILC closely matches external source statistics of the middle age cohort, while it underestimates the youngest cohort and overestimates the older group by a smaller proportion than SHIW. As for SHIW discrepancies can be worryingly large (e.g. -17.5% for youngest and +31.8% for oldest group in the Center of Italy).

Table 4 Population Statistics: frequencies and their standard errors

Variable	External source Frequencies (a)	SHIW Frequencies (b)	SHIW Standard error	SHIW Diff. % (b/a-1)*100	IT-SILC Frequencies (c)	IT-SILC Standard error	IT-SILC Diff. % (c/a-1)*100
Age 0-19							
Italy	11,408,746	10,669,166	179,935	-6.48	10,760,931	83,853	-5.68
North West	2,811,590	2,530,113	96,815	-10.01	2,673,483	43,519	-4.91
North East	2,069,019	2,164,051	72,273	4.59	1,958,606	28,342	-5.34
Center	2,106,040	1,737,163	75,409	-17.52	1,955,060	32,628	-7.17
South	4,422,097	4,237,839	109,069	-4.17	4,173,782	53,824	-5.62
Age 20-65							
Italy	37,179,179	36,436,238	337,064	-2.00	36,948,154	151,911	-0.62
North West	9,872,235	9,598,375	194,968	-2.77	9,760,948	80,082	-1.13
North East	7,124,107	7,241,443	134,291	1.65	7,039,530	51,288	-1.19
Center	7,293,877	6,809,238	147,241	-6.64	7,254,419	62,513	-0.54
South	12,888,960	12,787,182	186,642	-0.79	12,893,257	95,458	0.03
Age > 65							
Italy	11,457,143	12,616,590	185,011	10.12	12,017,396	70,712	4.89
North West	3,233,551	3,704,250	107,748	14.56	3,387,243	39,170	4.75
North East	2,279,994	1,946,118	53,937	-14.64	2,389,457	25,007	4.80
Center	2,398,411	3,161,605	98,630	31.82	2,511,477	29,813	4.71
South	3,545,187	3,804,617	98,284	7.32	3,729,219	40,583	5.19

Source: our elaboration on DemoIstat, values as 1st January 2009 (External Source) and SHIW and IT-SILC data.
Note: Following EUROMOD conventions, in IT-SILC individuals who were born after the income reference period have been dropped (see Ceriani et al., 2012).

With respect to the education levels, both datasets underestimate population with at most elementary education and lower secondary education (see Table 5). Upper secondary education is slightly overestimated in both datasets (by 5% in SHIW and by 4% in IT-SILC), while tertiary education is well approximated by SHIW, and overestimated by IT-SILC (4%). IT-SILC better fits education distribution by geographical area for all education categories but lower secondary education.

Table 5 Population by highest level of education attained: frequencies and their standard errors

Variable	External source Frequencies (a)	SHIW Frequencies (b)	SHIW Standard error	Diff. % (b/a-1)*100	IT-SILC Frequencies (c)	IT-SILC Standard error	Diff. % (c/a-1)*100
Elementary Education							
Italy	12,379,000	11,980,676	174,686	-3.22	11,576,464	73,665	-6.48
North West	2,980,000	2,740,537	95,718	-8.04	2,854,287	37,062	-4.22
North East	2,323,000	1,748,543	51,924	-24.73	2,234,921	24,530	-3.79
Center	2,271,000	2,533,318	79,516	11.55	2,226,255	28,632	-1.97
South	4,805,000	4,958,278	110,357	3.19	4,261,001	48,345	-11.32
Lower Secondary							
Italy	16,285,000	15,950,263	215,876	-2.06	15,690,681	99,741	-3.65
North West	4,393,000	4,307,785	130,857	-1.94	4,276,672	54,038	-2.65
North East	2,976,000	2,990,028	76,817	0.47	2,831,499	32,224	-4.86
Center	2,925,000	2,843,163	86,899	-2.80	2,786,859	38,068	-4.72
South	5,991,000	5,809,287	125,616	-3.03	5,795,651	63,650	-3.26
Upper Secondary							
Italy	17,077,000	17,915,782	238,861	4.91	17,830,531	101,128	4.41
North West	4,751,000	5,009,707	135,033	5.45	4,886,090	53,287	2.84
North East	3,471,000	3,882,779	99,289	11.86	3,628,807	36,691	4.55
Center	3,666,000	3,529,855	107,904	-3.71	3,798,268	44,818	3.61
South	5,190,000	5,493,441	129,806	5.85	5,517,366	60,352	6.31
Tertiary							
Italy	5,574,000	5,566,073	140,786	-0.14	5,780,476	60,781	3.70
North West	1,574,000	1,814,768	81,955	15.30	1,654,332	34,060	5.10
North East	1,053,000	983,817	57,247	-6.57	1,087,134	19,946	3.24
Center	1,297,000	1,421,667	80,259	9.61	1,305,711	25,910	0.67
South	1,650,000	1,345,821	52,594	-18.44	1,733,299	35,096	5.05

Source: our elaboration on Istat (2010d), SHIW and IT-SILC.

Note: Elementary education include the set of population who did not complete any education level or completed just the elementary school (grade 1 to 5). Lower secondary education is the group of people whose highest level of education is grade eight. Upper secondary education refers to people who received a diploma and Tertiary is the set of individuals who attain an undergraduate or graduate degree. Population is age 15 and more in the External Source and in SHIW, while it is age 16 and more in IT-SILC due to data availability

Self-employed individuals and pensioners are underestimated in SHIW and overestimated in IT-SILC (see Table 6). Self-employed individuals are underestimated in SHIW by 11% and overestimated in IT-SILC by 10%. Pensioners are underestimated in SHIW by 3% and overestimates in IT-SILC by 4%. Finally, employees are underestimated in SHIW (by 13%) and overestimated in IT-SILC (by 4%). Both datasets are not reliable in terms of the geographical distribution of employment; self-employment and pension income (see Table 6).

Table 6 Population Statistics: frequencies and their standard errors

Variable	External source Frequencies (a)	SHIW Frequencies (b)	SHIW Standard error	Diff. % (b/a-1)*100	IT-SILC Frequencies (c)	IT-SILC Standard error	Diff. % (c/a-1)*100
Employees							
Italy	21,611,778	18,734,187	228,015	-13.31	22,485,222	118,991	4.04
North West	5,985,582	5,266,597	133,671	-12.01	6,325,447	64,905	5.68
North East	4,637,931	4,296,697	95,571	-7.36	4,273,878	43,170	-7.85
Center	3,944,874	3,752,849	104,983	-4.87	4,627,595	48,960	17.31
South	5,734,778	5,418,044	116,004	-5.52	7,258,302	70,304	26.57
Self-employed							
Italy	6,117,343	5,438,250	155,809	-11.10	6,717,910	71,173	9.82
North West	2,067,939	1,712,138	98,638	-17.21	1,840,272	40,080	-11.01
North East	1,280,320	1,217,732	65,355	-4.89	1,308,766	24,974	2.22
Center	1,488,679	1,045,498	68,217	-29.77	1,393,713	26,769	-6.38
South	2,128,093	1,462,882	68,435	-31.26	2,175,159	43,040	2.21
Pensioners							
Italy	15,323,148	14,844,649	198,842	-3.12	15,877,444	86,908	3.62
North West	4,439,474	4,671,715	122,667	5.23	4,531,924	47,265	2.08
North East	2,864,306	2,504,802	64,410	-12.55	2,957,492	29,571	3.25
Center	3,055,415	3,412,963	97,978	11.70	3,232,561	36,895	5.80
South	4,476,872	4,255,169	101,464	-4.95	5,155,467	51,590	15.16

Source: our elaboration on MEF (2013) (external source) and SHIW and IT-SILC data.

Note: We count among employees all those who receive employment income and similarly for self-employment and pensioners, consistently with the external sources. Notice that the sum of frequencies in the external sources by geographic partition does not sum up to the total. This is because there is income which cannot be assigned to any Italian region.

As for the sector of activity there is no clear advantage of a dataset over the other. The distribution of the population by sector of activity shows that both datasets underestimate the share of population employed in agriculture (by 32% SHIW and by 14% IT-SILC) and in services (by roughly 10% both datasets) and overestimate the share of population employed in the industrial sector (see Table 7).

Finally, one should also notice that estimated standard errors are smaller for all frequencies estimated using IT-SILC compared with those estimated using SHIW, most likely as a consequence of a larger sample size.

Table 7 Population by sector of activity: frequencies and their standard errors

Variable	External source Frequencies (a)	SHIW Frequencies (b)	SHIW Standard error	Diff. % (b/a-1)*100	IT-SILC Frequencies (c)	IT-SILC Standard error	Diff. % (c/a-1)*100
Agriculture	1,287,100	878,337	44,619	-31.76	1,107,662	30,118	-13.94
Industry	6,988,500	7,486,766	150,641	7.13	7,145,046	70,310	2.24
Services	16,662,900	14,908,539	231,390	-10.53	15,034,130	102,976	-9.77

Source: our elaboration on Istat, Employment breakdown by industry (NACE Rev.2) - annual national data, 2009, wired at: http://dati.istat.it/? lang=en

4.3. Net incomes and income distribution

As there exists no external source providing population estimates of net incomes, we can only compare statistics on net incomes between the two surveys without any external validation. Table 8 compares the two datasets in terms of the main net income sources: employment, self-employment and pension incomes.

The average pension income and income from self-employment are higher in SHIW than in IT-SILC (by 1.5% in the case of pension income and by 12% for self-employment income), while both datasets give almost the same picture in terms of average net employment income. When splitting the Italian population into four macro regions, we note a huge variability in the discrepancies between the two datasets.

Table 8 Yearly Net Incomes: averages and their standard errors

	SHIW		IT-SILC		Diff %
	euro (b)	Std. Err.	euro (c)	Std. Err.	(c/b-1)*100
Employment Income					
Italy	15,950	2.17	15,887	2.13	-0.40
North West	17,633	4.63	17,277	4.24	-2.02
North East	15,724	3.82	16,212	4.32	3.11
Center	16,998	5.02	16,177	4.90	-4.83
South	13,759	3.66	14,115	3.60	2.59
Self-Employment Income					
Italy	21,109	8.16	18,685	7.84	-11.48
North West	23,018	17.53	20,848	16.22	-9.43
North East	24,151	20.28	21,052	20.66	-12.83
Center	20,043	13.94	19,447	17.58	-2.98
South	17,105	9.78	14,527	9.00	-15.07
Pension Income					
Italy	12,354	2.04	12,164	2.53	-1.54
North West	12,807	3.24	12,301	4.88	-3.95
North East	12,450	8.39	11,877	6.75	-4.60
Center	13,210	2.37	12,831	5.14	-2.87
South	10,881	2.11	11,708	2.95	7.61

Source: our elaboration on SHIW and IT-SILC data.

Using SHIW poverty and inequality figures are higher than using IT-SILC.[15] Results are summarized in Table 9, where we show headcount index and Gini inequality index. All indices are computed using household equivalent individual income (i.e. each individual is given his total household disposable income divided by the squared root of his family size). The poverty line is set to 60% of median household equivalent disposable income (and it is equal to 10,797 euro in IT-SILC and to 9,246 euro in SHIW). In IT-SILC the percentage of poor individuals is 19%, whereas in SHIW it is 21%. Gini index is 0.32 according to estimates based on IT-SILC and 0.35 if using SHIW.

Table 9 Yearly Net Incomes: poverty index, inequality index and their standard errors

	SHIW		IT-SILC		Diff %
	index (b)	Std. Err.	index (c)	Std. Err.	(c/b-1)*100
Poverty					
Headcount index	20.53	0.005	19.34	0.004	-5.82
Inequality					
Gini Index	35.00	0.006	31.60	0.003	-9.81

Source: our elaboration on SHIW and IT-SILC data.
Note 1: Poverty and Gini inequality computed on total household disposable income divided by the square root of household size

As a final comment to this section, we conclude, similarly to Ciani and Fresu (2011), that IT-SILC seems to be a rich source of data which offers more advantages than SHIW for microsimulation modelling, mainly due to (i) the larger sample size, (ii) the integration of net income information provided by interviewees with administrative records, (iii) the availability of additional and more disaggregated income variables than in SHIW and (iv) the smaller size of the standard error for most statistics considered and of interest for tax-benefit microsimulation. On the other hand, SHIW has a larger informational content than IT-SILC about the level of (i) financial and (ii) real estate wealth owned by the households, but the former does not enter the tax base for the Italian PIT and the latter could be used to integrate IT-SILC dataset as we will suggest below.

5. VALIDATION OF TABEITA OUTPUT

In this section we provide a thorough validation exercise of the gross income output simulated by TABEITA using SHIW and IT-SILC input data, contrasting it with external administrative sources, which are regarded as the benchmark to assess the reliability of microsimulation models (Subsection 5.1). Then, we will compare gross income provided in IT-SILC (produced using exact matching with administrative sources and Siena microsimulation model) and gross incomes simulated using only disposable income declared by survey respondents and the use of the TABEITA model (Subsection 5.2). Finally, we put forward a way to get the best for microsimulation analysis for Italy out of all available data (Subsection 5.3).

5.1. Gross Income Variables

In this section we compare gross incomes simulated by running TABEITA on SHIW and on IT-SILC (henceforth $SHIW_T$ and $IT\text{-}SILC_T$ respectively, where the subscript means that they have been obtained by using TABEITA net-to-gross routine). Table 10 summarizes the results, with and without calibration for tax evasion. We report frequencies and amount in euro for the external source (tax records data provided by MEF, 2013), and the percentage difference from

the external source of $SHIW_T$ and IT-$SILC_T$ for the main components of Irpef, the Italian PIT.[16]

Table 10 shows that the frequency of total taxable income units is close to the external source in IT-$SILC_T$ and slightly overestimated in $SHIW_T$. $SHIW_T$ underestimates frequencies and amount of employment income (-13% and -15%, respectively), while the figures for IT-$SILC_T$ are closer to the true values (+4% and +6%).

With respect to self-employment, without assuming tax evasion, as opposed to external statistics $SHIW_T$ shows fewer but richer self-employed taxpayers (frequencies -11%, amounts +30%); on the other hand, both frequencies and average amounts are overestimated in IT-$SILC_T$ (+10%, +45%, respectively). As expected, assuming tax evasion for self-employment income, both $SHIW_T$ and IT-$SILC_T$ average income estimates get closer to the external source.[17]

Income from pension is slightly underestimated with SHIW and overestimated with IT-SILC, in both cases with differences from external sources not larger than 6%.

Total deductions frequencies are underestimated (more in $SHIW_T$ than in IT-$SILC_T$), while amounts become closer to external sources when taking evasion into account (mainly due to self-employment social contributions, which are a major component of tax deductions).

Both datasets largely underestimate the number of home units, with IT-SILC underreporting by 87% the number of real estate and other buildings. This is due to the fact that IT-SILC provides no information about non-rented properties.

Overall, the main difference between $SHIW_T$ and IT-$SILC_T$ gross income variables concerns employment and self-employment income, which are underestimated in $SHIW_T$ and is reflected in a large underestimation of net regional and national Irpef.

Table 10 Comparison between External Sources and TABEITA (SHIWT and IT-SILCT) gross incomes

Variable	External Source		No evasion				Evasion			
			SHIW$_T$		IT-SILC$_T$		SHIW$_T$		IT-SILC$_T$	
	Frequencies	Amounts euro	Freq Diff%	Amount Diff%	Freq Diff%	Amount Diff%	Freq Diff%	Amount Diff%	Freq Diff%	Amount Diff%
Total taxable income	41,466,397	782,593,452	3.54	-4.47	-0.72	9.64	3.54	-8.69	-0.72	3.23
Employment income	21,611,778	418,740,720	-13.31	-14.61	4.04	5.94	-13.31	-14.67	4.04	5.72
Self-employment income	6,117,343	109,565,087	-11.10	30.30	9.82	44.63	-11.10	0.96	9.82	0.34
Pension income	15,323,148	213,594,560	-3.12	-1.75	3.62	6.53	-3.12	-1.96	3.62	6.28
Total deductions	12,687,840	21,721,425	-44.31	30.25	-22.60	43.59	-44.55	5.74	-22.05	7.10
Net taxable income	40,249,514	753,556,569	5.19	-5.81	2.16	8.43	5.17	-9.48	2.16	2.82
Total tax credit	39,423,594	62,917,813	-6.00	1.04	2.10	0.67	-5.67	2.55	2.49	2.40
Net personal income tax	31,087,681	146,157,039	6.32	-11.54	7.57	9.20	5.39	-18.17	5.59	-0.85
Regional additional income tax	30,652,846	8,633,217	7.83	-9.20	9.10	7.87	6.88	-13.46	7.09	1.13
Cadastral income main residence	29,776,305	10,551,000	-18.70	0.00	-21.19	0.00	-18.70	0.00	-21.19	0.00
Cadastral income other buildings	17,513,880	7,723,079	-14.19	2.69	-86.97	-55.97	-14.19	2.69	-86.97	-52.71

Source. our elaboration on MEF (2013) (External Source) and SHIW and IT-SILC data.

Note: Freq Diff % = frequencies SHIWT (IT-SILCT) over external source frequencies; Amount Diff % = amount in euro SHIWT (IT-SILCT) over external source amounts. The net personal income tax is the due tax, net of all tax credits. The net taxable income is the tax base, i.e. gross taxable income net of tax allowances.

5.2. Comparing TABEITA gross incomes and IT-SILC gross incomes

Starting from the 2007 IT-SILC survey, the Italian National Statistical Institute (Istat) has been providing also gross incomes together with the net amounts (Istat, 2011). This novelty makes IT-SILC the first national and official database in Italy that includes gross income variables.

Information available in the administrative sources are used for cleaning the original dataset and controlling for possible incoherencies.

Gross incomes are obtained as a sum of: the net incomes declared in the survey, income tax and employees and employers social contributions obtained from administrative sources,[18] if available, or otherwise simulated using the Siena microsimulation model.

The Siena microsimulation model uses the original IT-SILC dataset including net incomes, after being cleaned and prepared, as input file for simulating personal income taxes and social insurance contributions. The gross income tax variables simulated by the model are then compared with external administrative data. If the administrative information on income tax is available for a record, the simulated tax is replaced with the value corresponding to the administrative source, the latter being considered as the true amount effectively paid by the the

taxpayer. Therefore, the main contribution of the microsimulation model is to provide an estimation of income tax and social contributions not included in the administrative source data.

The personal income tax considered in IT-SILC for employment incomes, pensions and unemployment incomes is equal to the withholding tax. For the self-employment income, the personal income tax includes the regional tax on business (Irap) computed by applying the regional Irap tax rate to the self-employment income included in the total taxable income.[19]

The IT-SILC three gross income categories (employment income, self-employment income and pension income) are obtained as aggregation of income variables different from what done in MEF. In particular, in IT-SILC gross self-employment income is the sum of self employment income, maternity benefits for self-employed and incomes from temporary contract, whereas MEF does not include income from temporary contracts. With respect to employment income, IT-SILC generates gross employment income as sum of employment salary, employment additional compensation, other compensations (profit sharing and bonuses), additional payments based on productivity, income from arrears, and additional months salary. MEF includes in this category also income from temporary jobs, unemployment benefits (for vocational training, redundancy payments, unemployment, mobility or early pension allowances), and private transfers including maintenance payments. Finally, gross pension income in IT-SILC is the sum of old age pensions, survivor' pensions, disability and invalidity pensions, social pensions. In MEF, instead, social and invalidity pensions are not included as pensions as they are tax exempt and treated as social assistance programs. This implies that, except for total taxable income no validation of IT-SILC data is possible. However, as TABEITA aggregate statistics are built consistently with those published by MEF and the validation exercise reported in Table 10 is rather satisfactory, one can indirectly assess the validity of IT-SILC gross income data using IT-$SILC_T$ with evasion, upon appropriate definition of income sources to make it comparable with the three gross income categories provided in IT-SILC. Table 11 shows that although the difference in total amounts is limited as for pension and employment income, it is large as for self-employment income. In fact, IT-SILC total gross self-employment income is more than 30% larger than what produced with IT-$SILC_T$, which is roughly consistent with tax records regarding Irpef and produced by MEF. This is also reflected in a larger gross total taxable income in IT-SILC. Part of this discrepancy could be due to the fact that IT-SILC provides gross self-employment income including also the regional tax on business (Irap), although IT-SILC does not provide information to properly assess this. It however remains that using IT-SILC data instead of IT-$SILC_T$ one would most likely largely overestimate the amount of taxable income

received and tax paid by self-employed taxpayers.

Table 11 Comparisons between External Sources and TABEITA (IT-SILCT) and IT-SILC gross incomes

Variable	External source		$IT\text{-}SILC_T$		IT-SILC	
	Frequencies	Amounts million euro	Frequencies	Amounts million euro	Frequencies	Amounts million euro
Total taxable income	41,466,397	782,593	40,993,736	795,174	41,048,992	848,645
Employment income	na	na	21,485,120	428,534	21,485,120	431,602
Self-employment income	na	na	7,498,613	129,802	7,498,613	171,175
Pension income	na	na	16,541,269	240,099	16,677,270	245,868

Source: our elaboration on MEF (2013) (External Source) and IT-SILC data.

Note: "na" stands for not available. In fact, IT-SILC definition of gross employment, self-employment and pension income is not consistent with what provided in MEF aggregate tables used for validation of TABEITA (Table 10). As there exist no external sources to validate IT-SILC aggregated variables except for total taxable income we aggregated gross variables simulated using TABEITA (IT-SILCT) consistently with IT-SILC definition of gross variables and indirectly validate IT-SILC data. IT-SILC gross incomes are released as aggregate variables subject to administrative exact matching that we are not able to totally replicate in TABEITA. This explains the small differences between $IT\text{-}SILC_T$ and IT-SILC frequencies for pension income.

5.3. How to make the best of two datasets

Our analysis of Table 11 suggests using disposable incomes as declared by respondents in interviews and simulating the individual personal income tax, with calibration for tax evasion. Moreover, our reading of Table 10 suggests us to prefer IT-SILC as opposed to SHIW, but for one income component, namely cadastral income on second homes and other buildings.

In fact, in both data sets we calibrate cadastral income of building and real estate using the self-assessed market value of homes by homeowners, and the actual or – when missing – the self-assessed rent obtained from second homes and other properties by owners. The main difference between SHIW and IT-SILC is that only the former provides details about imputed rents of not rented second homes and other properties. This explains why the average amount of cadastral income from other building using IT-SILCT is so much underestimated, as IT-SILCT sets this income equal zero lacking appropriate information in the input data set. One may think of ignoring this limitation of IT-SILCT as it accounts for a meagre 1% of total taxable income. However, it could be a large share of income for some households and it could be particularly relevant for specific tax-benefit applications such as those estimating the cost of taxation of real estate and building property income (see Avram et al., 2012).

Here we suggest a different approach. We match information on second homes and other not rented properties coming from SHIW in the IT-SILC data set and use it to produce an improved version of $IT\text{-}SILC_T$. In Table 12 we present validation results where we assigned to all

household units who declared to own other buildings without renting them a value of imputed rents equal to the average amounts received by households who are in the same condition and belong to the same percentile of equivalent income (defined as the total household income divided by the square-root of the household size) as the SHIW data set.[20] As shown in the last line of this table, the average amount of cadastral income from other buildings is close to external source[21] and the impact on the validation of other simulated variables is negligible.

Table 12 Comparisons between External Sources and TABEITA with calibration ($\mathrm{IT-SILC}_T^C$) of cadastral incomes for building different from principal residence

Variable	External Source		$\mathrm{IT-SILC}_T^C$			
			Evasion		No evasion	
	Frequencies	Amounts euro	Freq Diff %	Amount Diff %	Freq Diff %	Amount Diff %
Total taxable income	41,466,397	782,593,452	0.59	10.39	0.59	4.04
Employment income	21,611,778	418,740,720	4.04	6.10	4.04	5.87
Self-employment income	6,117,343	109,565,087	9.82	44.76	9.82	0.53
Pension income	15,323,148	213,594,560	3.62	6.62	3.62	6.39
Total deductions	12,687,840	21,721,425	-22.72	43.75	-22.07	7.20
Net taxable income	40,249,514	753,556,569	3.35	9.19	3.36	3.65
Total tax credit	39,423,594	62,917,813	2.12	0.38	2.36	2.05
Net personal income tax	31,087,681	146,157,039	9.27	10.55	7.39	0.59
Regional additional income tax	30,652,846	8,633,217	10.82	8.82	8.91	2.16
Cadastral income main residence	29,776,305	10,551,000	-21.19	0.00	-21.19	0.00
Cadastral income other buildings	17,513,880	7,723,079	-56.83	-6.65	-56.83	0.26

Source: our elaboration on MEF (2013) (External Source) and IT-SILC data.
Note: Freq Diff % = frequencies $\mathrm{IT-SILC}_T^C$ over external source frequencies; Amount Diff % = amount in euro $\mathrm{IT-SILC}_T^C$ over external source amounts.

6. CONCLUDING REMARKS

This paper aims at providing an analysis of pros and cons of alternative datasets for tax-benefit microsimulation for Italy. We analysed all available possibilities, namely using gross income data resulted from a consistent net-to-gross procedure as the one provided by the TABEITA microsimulation model (and choosing between either SHIW or IT-SILC data) or using the gross income data provided by IT-SILC based on Siena microsimulation models and exact matching with administrative data.

Our results suggest that IT-SILC improves in the regional representativeness of the Italian population and does not perform worse than SHIW as for most demographic characteristics. As for gross income variables, simulated gross income variables using IT-SILC income data allow the best fit with external tax statistics. Moreover, we show that valuable source of information about real estate properties found in SHIW can be exploited to improve the validation of

simulated gross income data using IT-SILC.

This paper is relevant for microsimulation models in general, whenever more than one data set is available. We suggest that it is important to first assess the quality of available data, and in particular issues such as sample design and non-response rate. Then, the set of available information in the alternative dataset has to be compared, in order to analyze the level of detail and to the sample representativeness. We moreover suggest that input data may benefit by matching information available in the different data sources and that the assessment of standard errors of main statistics produced can be very informative about the pros and cons of different data sets for microsimulation analysis.

ACKNOWLEDGMENTS

We are very grateful to Francesco Figari and Iva Tasseva for insightful discussions and suggestions, to Simone Pellegrino and participants to the 2012 EUROMODupdate project meeting in Bucharest and to three anonymous referees for much appreciated comments and advices.

REFERENCES

Andrigde R. R. and J. A. L. Little (2011) 'Proxy Pattern-Mixture Analysis for Survey Non-response', *Journal of Official Statistics*, 27(2): 153-180.

Atkinson A. B. and H. Sutherland (1983) 'Hypothetical families in the DHSS Tax/Benefit model and families in the Family Expenditure Survey 1980', SSRC Program on Taxation, Incentives and the Distribution of Income, Research Note No. 1. STICERD, London: School of Economics.

Avram S., F. Figari, C. Leventi, H. Levy, J. Navicke, M. Matsaganis, E. Militaru, A. Paulus, O. Rastrigina and H. Sutherland (2012) 'The distributional effects of fiscal consolidation in 9 EU countries', Social Situation Observatory Research Note 01/2012.

Bank of Italy (2010) 'Sample Surveys Household Income and Wealth in 2008', Supplements to the Statistical Bulletin, Volume XX - Number 8. Rome: Bank of Italy.

Bernasconi M. and A. Marenzi (1999) 'Gli effetti redistributivi dell'evasione fiscale in Italia' in *Ricerche quantitative per la politica economica*, pp. 1-38. Rome: Bank of Italy.

Betti G., Donatiello G. and Verma V. (2011) 'The Siena Microsimulation Model (SM2) for net-gross conversion of EU-SILC income variables', *The International Journal of Microsimulation* 4: 35-53.

Bollinger C. and M. David (1997) 'Modelling discrete choice with response error: Food Stamp participation', *Journal of the American Statistical Association*, 92:827-835.

Brandolini A. (2008) 'Income Inequality in Italy: Facts and Measurement', XLIV Riunione Scientifica, CLEUP, Padova.

Ceriani L., F. Figari and C. Gigliarano (2011) 'EUROMOD Country Report Italy (IT) 2006 - 2009', EUROMOD Country Reports.
https://www.iser.essex.ac.uk/files/euromod/country-reports/CR_IT2006-09_final_6-12-11.pdf

Ceriani L., F. Figari and C. Fiorio (2012) 'EUROMOD Country Report Italy (IT) 2007 - 2010', EUROMOD Country Reports.

https://www.iser.essex.ac.uk/files/euromod/country-reports/year-3/CR_IT_2007-2010_Y3_FINAL.pdf

Ciani E. and D. Fresu (2011) 'From SHIW to IT-SILC: construction and representativeness of the new CAPP_DYN first-year population', CAPPaper n. 92.

Coulter F., Heady, C., Lawson, C., Smith, S., and G. Stark (1998) 'A Microsimulation Model of Personal Tax and Social Security Benefits in the Czech Republic' in Spahn, P. and M. Pearson (Eds.), *Tax Modelling for Economies in Transition,* Macmillan Press Ltd, 163-189.

Eurostat (2007) *Comparative EU statistics on Income and Living Conditions: Issues and Challenges*, Proceedings of the EU-SILC conference, Helsinki, 6-8 November 2006.

D'Alessio G and I. Faiella (2002) 'Non-response behaviour in ths Bank of Italy's Survey of Household Income and Wealth', Temi di Discussione del Servizio Studi, N. 462, Bank of Italy.

Fiorio C.V. and F. D'Amuri (2005) 'Workers' tax evasion', *Giornale degli Economisti ed Annali d'Economia,* 118(64), N° 2/3: 247-270.

Fiorio C.V. (2009) *Microsimulation and analysis of income distribution*, VDM Verlag.

Groves R. M. and E. Peytcheva (2008) 'The impact of non-response rates on non-response bias', *Public Opinion Quarterly*, 72(2): 167-189.

Gupta, A. and V Kapur (2000) *Microsimulation in Government Policy and Forecasting*. Contributions to economic analysis, Amsterdam: North-Holland, Amsterdam.

Hernandez M., Pudney, S. and R. Hancock (2007) 'The welfare cost of means-testing: pensioner participation in income support', *Journal of Applied Econometrics*, 22(3): 581–598.

Iacovou M., Kaminska O. and H. Levy (2012) 'Using EU-SILC data for cross-national analysis: strengths, problems and recommendations', ISER Working Paper Series No. 2012-03, Institute for Social and Economic Research, University of Essex.

Istat (2008a) *L'indagine europea sui redditi e le condizioni di vita delle famiglie (EU-SILC)*, Metodi e Norme n. 37, Rome: Istat.

Istat (2008b) *I beneficiari delle prestazioni pensionistiche, Anno 2008*, Rome: Istat.

Istat (2009) *Integrazione di dati campionari EU-SILC con dati di fonte amministrativa*, Metodi e Norme n. 38, Rome: Istat.

Istat (2010a) *La misura dell'economia sommersa secondo le statistiche ufficiali Anni 2000-2008*, Rome: Istat. http://www.istat.it/it/archivio/4384.

Istat (2010b) *Intermediate quality report cross-sectional survey 2008 Italy*, available at http://epp.eurostat.ec.europa.eu/portal/page/portal/income_social_inclusion_living_conditions/documents/tab9/2008\20Intermediate\%20Quality\%20Report\%20IT.pdf

Istat (2010c) *Distribuzione del reddito e condizioni di vita in Italia Anni 2008-2009*, Statistiche in breve, Rome: Istat.

Istat (2010d) *Forze Lavoro, Media 2001,* Annuari, Rome: Istat.

Istat (2011) *La metodologia di stima dei redditi lordi nell'indagine EU-SILC*, Metodi e Norme n. 49, Rome: Istat.

Juhàsz I. (1998) The Hungarian Personal Income Tax Model in Spahn, P. and M. Pearson (Eds.), *Tax Modelling for Economies in Transition,* Macmillan Press Ltd, 191-205.

Lelkes O. (2007) 'Tax Benefit Microsimulation Models in Eastern Europe', *The International Journal of Microsimulation,* 1(1): 54-56.

Lelkes O. and H. Sutherland (2009) *Tax and Benefits Policies in the Enlarged Europe: Assessing the impact with microsimulation models*, Farnham: Ashgate.

Mantovani D. (1998), *Manuale DIRIMOD95*, mimeo, Prometeia, Bologna.

MEF (2008) *Analisi Statistiche - Dichiarazioni 2009 - Anno d'Imposta 2008*, Department of Finance. www.finanze.gov.it.

Mitton L., Sutherland, H., and M. Weeks (2000) *Microsimulation Modelling for Policy Analysis: Challenges and Innovations*, Cambridge: Cambridge University Press.

OECD (2010) *Taxing Wages*, Paris: OECD.

Pissarides C. A. and G. Weber (1989) 'An expenditure-based estimate of Britain's black economy', *Journal of Public Economics* 39(1): 17-32.

Savegnago M. (2008) 'Rappresentatività campionaria delle indagini Banca d'Italia e SILC', CAPPaper n. 46.

Schouten B., Cobben F. and J. Bethlehem (2009) 'Indicators for the representativeness of survey response', *Survey Methodology*, 35(1): 101-113.

Statistics Canada (2003) *Quality Guidelines*, Fourth Edition, available at http://www.statcan.gc.ca/pub/12-539-x/4147797-eng.htm

Sutherland H. (1991) 'Constructing a Tax-Benefit Model: What Advice Can One Give? ', *Review of Income and Wealth*, 37(2): 199-219.

Sutherland H. (1995) 'Static Microsimulation Models in Europe: a Survey', Department of Applied Economics, University of Cambridge, Cambridge Working Papers in Economics.

Sutherland H. and F. Figari (2013) 'EUROMOD: the European Union tax-benefit microsimulation model', *International Journal of Microsimulation*, 6(1): 4-26.

Wagner J. (2010) 'The fraction of missing information as a tool for monitoring the quality of survey data', *Public Opinion Quarterly*, 74: 223-243.

7. APPENDIX

Tables in Appendix show average values and standard errors of figures in Tables 10 and 12.

Table 13 TABEITA ($SHIW_T$ and $IT\text{-}SILC_T$) gross incomes. Averages and their standard errors

Variable	No evasion				Evasion			
	$SHIW_T$		$IT\text{-}SILC_T$		$SHIW_T$		$IT\text{-}SILC_T$	
	Average euro	Std. Err.	Average euro	Std. Err.	Average euro	Std. Err.	Average euro	Std. Err.
Total taxable income	17,413.61	176.07	20,842.63	113.96	16,644.55	162.32	19,624.06	99.61
Employment income	19,086.86	184.06	19,728.63	119.78	19,072.34	183.68	19,687.35	119.43
Self-employment income	26,251.73	831.85	23,588.21	400.25	20,341.09	648.51	16,365.29	280.38
Pension income	14,136.62	205.80	14,331.29	93.27	14,106.22	206.52	14,297.66	92.81
Total deductions	4,003.99	82.94	3,176.11	40.75	3,264.82	69.93	2,352.12	30.27
Net taxable income	16,764.81	170.18	19,870.83	106.74	16,113.46	158.51	18,843.63	95.18
Total tax credit	1,715.53	7.12	1,573.64	3.97	1,735.05	7.16	1,594.42	3.95
Net personal income tax	3,911.55	79.73	4,772.70	47.73	3,650.65	73.93	4,414.69	41.75
Regional additional income tax	237.15	2.80	278.49	1.69	228.04	2.61	265.96	1.51
Cadastral income main residence	435.84	3.68	449.60	2.82	435.84	3.68	449.60	2.82
Cadastral income other buildings	527.71	20.24	1,490.07	39.08	527.71	20.24	1,600.44	41.98

Source: our elaboration on SHIW and IT-SILC data.
Note: Std. Err. refers to the standard error of the average.

Table 14 TABEITA with calibration ($\mathrm{IT}-\mathrm{SILC}_T^C$) of cadastral incomes for building different from principal residence. Averages and their standard errors

Variable	$\mathrm{IT}-\mathrm{SILC}_T^C$			
	No evasion		Evasion	
	Average Euro	Std. Err.	Average Euro	Std. Err.
Total taxable income	20,712.56	113.76	19,520.17	99.66
Employment income	19,758.79	119.96	19,716.68	119.53
Self-employment income	23,609.04	400.31	16,395.09	280.91
Pension income	14,343.90	93.15	14,312.55	92.81
Total deductions	3,184.75	40.78	2,355.07	30.36
Net taxable income	19,780.29	106.7	18,774.24	95.33
Total tax credit	1,568.60	3.95	1,591.00	3.96
Net personal income tax	4,756.20	47.40	4,404.03	41.51
Regional additional income tax	276.55	1.68	264.19	1.51
Cadastral income main residence	449.60	2.82	449.60	2.82
Cadastral income other buildings	953.51	13.88	1,024.14	14.91

Source: our elaboration on SHIW and IT-SILC data.
Note: Std. Err. refers to the standard error of the average.

[1] Extensive comparisons between IT-SILC and SHIW data are also provided in Savegnago, 2008; Ciani and Fresu, 2011.

[2] In particular, when there is an inconsistency between the two income sources, IT-SILC records the maximum value. The explanation proposed to justify this rule is that if the highest value is the one from administrative records the respondent is underreporting in the survey. On the other hand, if the sample value is higher than the administrative one, the respondent is underreporting in the administrative records (such as tax records).

[3] The large difference between SHIW and IT-SILC response rate might also be explained by the fact that sampled households in IT-SILC are obliged by law to participate to the survey. However, Istat prefers not to underline the mandatory nature of the survey, but rather to encourage voluntary participation.

[4] The NUTS classification (Nomenclature of territorial units for statistics) has been created by Eurostat and it is a hierarchical system for dividing up the economic territory of the EU. NUTS II-level stands for the basic regions for the application of regional policies. For further details, see *http://epp.eurostat.ec.europa.eu/portal/page/portal/nuts_nomenclature/introduction*

[5] Although the analysis of benefits take-up is a relevant issue in some countries (e.g., see Bollinger and David, 1997, Hernandez et al., 2007), in Italy tax credits and tax deductions are claimed in the tax form involving no stigma or psychological dependency.

[6] We select the head of the unit as the individual with the largest taxable income, in case of draws the one who is older, in case of draws the man. In the few cases where no head is eventually identified, we select one randomly.

[7] Tax rules for 2008 Irpef define as independent for tax purposes all citizens earning more than euro 2,840.51.

[8] Alternatively, one could simulate withheld taxation by income source, but tax differences are on average small and we believe our approach is closer to the definition of disposable income provided in SHIW and IT-SILC.

[9] In particular this is done for tax deductions such as expenditures for disabled relatives, gifts to religious institutions, social security contributions paid to domestic collaborators and for tax credit for health care expenditures, for guide dogs, funeral expenditures, gifts to not-for-profit organisations, life insurance. Overall, all these items account for less than 0.9% of total taxable income.

[10] For details, refer to Ceriani at al., 2012.

[11] In TABEITA iterations are stopped when the left hand side of the direct tax function (2) using the simulated BT income y_S^B provides an AT income which is equal to y^A up to one euro.

[12] In the current version of TABEITA using SHIW we have not simulated family allowances and simply used the aggregate variable to define taxable employment income. This choice was also justified by the fact that our validation exercise on SHIW data (more details in Section 5) found no overestimation of employment income.

[13] TABEITA is programmed to use all available variables that enter the Italian PIT base and when one is missing in the input dataset used, it is replaced with a zero value.

[14] These discrepancies may be also due to the fact that the external sources includes also the institutionalised population which is not in the universe of interest of both surveys. However, due to data limitation, we are not able to separate these individuals in the external sources.

[15] This is consistent with the official estimates, see Istat (2010c) and Bank of Italy (2010).

[16] For average values and standard errors corresponding to the figures in Table 10, see Table 13 in the Appendix.

[17] The calibration coefficients are equal to 27% and 20% using IT-SILC$_T$ and SHIW$_T$ data, respectively.

[18] These include the MEF and the National Institutes for Social Security, namely Inps and Inpdap.

[19] TABEITA does not take into account the regional tax on business, as it concentrates only on direct taxes (Irap is formally a tax on production, not on incomes) and social contribution. The average Irap paid by self-employed individuals in 2009 was on average 1,310 euro (MEF, 2013).

[20] For average values and standard errors corresponding to the figures in Table 12, see Table 14 in the Appendix.

[21] The correct number of taxpayers with positive cadastral income is more difficult to assess, mostly due to the fact that we have no or little information on ownership shares. However, this does not affect the possibility of simulating the Property Tax, as this does not depend on personal income but only on the cadastral value of properties.