Università Commerciale Luigi Bocconi
Econpubblica
Centre for Research on the Public Sector

WORKING PAPER SERIES

Grossing-up and validation issues in an Italian tax-benefit microsimulation model

Francesco D'Amuri, Carlo V. Fiorio

Working Paper n. 117

December 2006

Grossing-up and validation issues in an Italian tax-benefit microsimulation model*

Francesco D'Amuri Carlo V. Fiorio

December 2006

Abstract

The aim of this paper is to give a detailed description of TABEITA, the tax-benefit Micro Simulation Model (MSM) developed and mantained at Econpubblica (Center for Research on the Public Sector, Bocconi University) for the analysis of the redistributive impact of the Italian Personal Income Tax. After a brief general introduction on the scope and main features of MSMs, all the main characteristics of TABEITA are discussed. Particular emphasis is posed on issues that are crucial for the reliability of a MSM: the grossing-up procedure and the validation of model outputs. A sample simulation of the 2004 and 2005 PIT reforms are reported in order to show one of the possible applications of the MSM. A concise guide for TABEITA users is provided in the appendix.

JEL codes: C42, C88, H24, H26.
Keywords: Tax-Benefit Microsimulation, Personal Income Taxation, Survey Methods.

*This article presents some results of an ongoing teamwork at Econpubblica by Marco Cavalli, Francesco D'Amuri and Carlo Fiorio. Sections 2, 2.1, 3, 4 and 7 were written by Carlo Fiorio, the remaining ones by Francesco D'Amuri and Carlo Fiorio. Financial support by MIUR (PRIN 2004, prot. 2004138335) is gratefully acknowledged. We also wish to thank Roberto Artoni and Alberto Zanardi for their support.

1 Introduction

In recent years, there has been an extensive development of simulation models for quantitative research in economics. The main aim of Micro Simulation Models (MSMs), based on individual or household data, is to analyze the impact of policy changes on the distribution of some target variables rather than on their mean, as it happens using regression techniques. With a MSM the immediate distributional impact of fiscal policies, such as an increase in child benefits, in income tax rates or in the minimum wage, can be modeled, and estimates of the characteristics of winners and losers and total cost can be computed. MSMs can also be used to project into the future and to assess the socio-economic consequences of an ageing population, or of changes in educational structure and in marriage patterns.

On the basis of the method employed to project the demographic characteristics of the sample in the future, MSMs can be classified as static or dynamic. In static MSMs, the projection of the sample demographic characteristics can be obtained through the variation of the sample weights, leaving unchanged the observations. In dynamic models, projection is obtained through the use of transition tables that modify the characteristics and the composition of the sample through time. Moreover, MSMs can be equipped with the estimates of the behavioral responses of individuals to simulated changes in economic policies (behavioral MSM).

The purpose of this article is to describe in detail TABEITA, the MSM currently developed and mantained at Econpubblica (Center for Research on the Public Sector, Bocconi University). TABEITA is a static MSM model and does not produce estimates of behavioral responses: it provides a detailed picture of the redistributive effects of fiscal reforms, without taking in consideration

the change in the economic behavior that could be induced by them. At the moment, stata modules have been computed in order to use the last 4 waves of the Bank of Italy's Survey of Households Income and Wealth (SHIW), covering the period 1998-2004. The first version of TABEITA (TABEITA98, using year 1998 SHIW data) was developed by Carlo Fiorio (Fiorio, 2004). Subsequent updates were developed by Francesco D'Amuri (TABEITA, using year 2000 SHIW data) and Marco Cavalli (TABEITA02 and TABEITA04, years 2002 and 2004). In this paper we focus mainly on the first two versions of the model. The main distinguishing features of TABEITA02 and TABEITA04 are described in Cavalli and Fiorio (2006). Section 2 discusses the data set employed and its main limitations, Sections 3 and 4 briefly outline the structure of the model, while Sections 5 and 6 discuss critical issues for the reliability of a MSM: grossing up and validation. The equivalence scale employed is described in Section 7, while Section 8 describes a simple exercise using TABEITA00.

2 The data set

Static models are generally based on sample surveys, which provide detailed information about individual and family characteristics, labor force status, housing status, earnings. The data set used by TABEITA, is the Survey of Household Income and Wealth (SHIW) published by the Bank of Italy. TABEITA actually employes the last four SHIW datasets (from year 1998 to 2004, the last available year). The SHIW is a long standing survey: it was started in the mid 1960s, was run about annually up to 1987, henceforth about every two years. The Bank of Italy has been paying particular attention to improve the quality of the data. For instance since 1995 an increasing number of interviews were performed using a computer to check consistency of answers

and particular attention was paid in formulating questions as clearly as possible with several trial interviews. At present the SHIWs are the main, if not the only, data set for Italian household MSMs and among the most frequently used for any kind of household income analysis at the national level in Italy (for a review of other data sets, see Brandolini, 1999). The sample is drawn in two stages (municipalities and households) with the stratification of the primary sampling units (municipalities) by units and size, to make it representative of the national population. Within each stratum, all municipalities with population of more than 40,000 were selected, while smaller towns were randomly included. Households were then selected randomly and a sampling weight, defined as the inverse of the probability of inclusion of each household in the sample, was attached to each observation. Data are checked before release: the strategy is either to drop the interview for the whole household if missing data cannot be reasonably inferred from other characteristics of the individual/household or to impute the missing data, often using regression models to forecast missing variables based on the personal characteristics of the individual/household involved. Data imputation is very low for most variables (Banca d'Italia, 2006, p.40).

2.1 Some limitations of the database

Among the limitations of the SHIW data set some affect any analysis of Italian household income, some others are specifically of interest for the reliability of MSMs. A first limitation of the dataset is the low rate of response. Participation in the survey is voluntary and not paid. Although all households were granted total anonymity, in 2004 only 36.4% of contacted households agreed to being interviewed. The low rate of response can cause a selectivity bias as some households seem to be more likely to refuse an interview. In fact, the

likelihood of accepting an interview decreases with increases in income, wealth and education of the household head, and the size of the town of residence (Banca d'Italia, 2006, p. 35). In order to mitigate the selectivity bias some measures are adopted, such as the replacement of refusing households with others from the same town. Some estimations of the selectivity bias on incomes recorded in SHIW show that the underestimation of household income is on average rather limited (Cannari and D'Alessio, 1992) estimate it at about 5%). Other limitations of this data set include the fact that the household is interviewed rather than the family. This leads to an overestimation of the average number of components, which cannot be corrected at all since the relevant information is missing. The interviews include only recall questions, i.e. questions referring to the previous year, reducing the precision of the reporting. An alternative approach would be to ask households to record all their incomes and expenditures of the coming week or month but it was discarded to keep a reasonable rate of response and to avoid approximations that come from extending the week or month to cover the whole year. Finally, data do not include information about people who do not have a registered dwelling or are in a hospital or other kind of institution. As for the limitations which are more relevant for MSMs, the main one refers to the type of income recorded: it refers to disposable income, excluding taxes and social contributions paid and benefit received. Hence, the first role of a MSM is to simulate the before-tax income before introducing any other policy simulation. This feature implies that, in contrast to other MSMs, no simulation error can be properly assessed (for the U.K. see Pudney and Sutherland, 1994).

2.2 Preliminary check of the dataset

In the SHIW dataset income variables are reported in separate form (i.e. in different files reporting different types of income) and in aggregate. Data imputation is performed by means of regression models. In order to avoid an excessive concentration around mean values, random drawings from a normal with mean zero and variance equal to that of the estimated residuals are added to the estimated values (Banca d'Italia (2006), p. 40).

Before using the SHIW dataset, a careful preliminary check is performed in order to assess the consistency of data with a particular focus on income variables (for details see D'Amuri (2004), pp. 71-80). A new dataset is created merging all the files reporting socioeconomic variables and compared to the aggregated dataset delivered by the Bank of Italy. In this way it is possible to check the consistency of the aggregation process and the extent and the outcomes of data imputation. No inconsistencies are present in the data (i.e. no errors are found in the aggregation of different variables), while the extent of data imputation is variable with the type of variable taken in consideration. In year 2000, a very low level of data imputation is found for employment income. The few cases of imputation (never more than 2% of the total population of perceivers of this type of income) are due to missreporting of employment benefits and of pension income. For self-employment income data imputation is higher (but still very low in relative terms, around 4% of the total of self-employed) and partly is due to imputation of investment for tax credits. The higher extent of data imputation reflects greater difficulties in the measurement of this type of income relative to employment income. Real estate and capital income are recorded at the family level, and are imputed to individuals using share of ownership, which is recorded for each individual in the household. Some discrepancy appears here, in particular as the reconstruction of

individual income starting from family income does not seem to strictly follow the imputation via ownership quotas, but no information is available on the Bank of Italy procedure for imputing real estate income. In the SHIW dataset used by TABEITA it was preferred to impute income at an individual level by following information on share of ownership.

3 The structure of the model

TABEITA[1] refers to the Italian Personal Income Taxation (PIT) net of social contributions. TABEITA can be described as a deterministic transformation of a given sample into a new one. Let $\mathbf{y^A}$ and $\mathbf{y^B}$ be the vectors of after-tax (AT) and before-tax (BT) income, respectively: the former vector is obtained from the latter through a tax transformation, say τ_i, $i = 1, 2..., N$, where N is the number of individuals in the sample. Since the data are net of taxes and social contributions, the first role of the model is to recover 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 individual base; the amount of tax each individual 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; there are several tax allowances which depend on a set of individual and family characteristics, such as the number of dependent children, whether the spouse is dependant, whether income comes from self-employment, employment or

[1]For an extended discussion of the model refer to Fiorio (2004).

pension, etc. Secondly, the tax transformation in (1) is highly non linear. This implies that y_B has to be obtained numerically, by recursive approximations. The tax transformation, τ_i, used to recover y^B is obtained from the tax code of the relevant year. Various assumptions about take-up rates of tax allowances could be introduced, however no uncertainty is considered here. Although the analysis of benefits take-up is a relevant issue in countries where welfare programs are widespread[2], in Italy there are no generalized unemployment benefit, income maintenance or house benefit schemes. The issue of non take-up is limited to tax allowances and tax deductions which do not involve issues of stigma or psychological dependency and is then less relevant. Moreover, this choice, together with the idea of not considering behavioral responses, allowed the model to be as simple and robust as possible.

In this model the main assumptions are that (a) the sample is representative of the population and contains enough details for simulation, (b) the tax and benefit legislation, τ_i, is perfectly known by the individual and applied without error. Although the first assumption is granted by the Bank of Italy who produced the data, the second is meant to keep variability to a minimum. An alternative solution would be to randomly include errors in the model assessing the relevance of such changes in the MSM. Here, instead, only systematic errors leading to under-reporting are considered and treated as tax evasion or tax avoidance; involuntary under-reporting is assumed to be off-setting with involuntary over-reporting errors. The probability of programming mistakes is kept to a minimum with a number of checks and a validation procedure.

TABEITA is developed in four different modules using Stata on a personal computer. The first module involves the preparation of the data base, and in particular the data consistency checks. The second is the grossing-up proce-

[2]For instance, see among others Fry and Stark (1993), Duclos (1995), Bollinger and David (1997), Pudney (2001) and Pudney et al. (2002).

dure. The third involves an estimation of tax evasion and the validation of the output. The last deals with the simulation of alternative scenarios.

The model allows one to compute the amount of tax allowances, of tax base deductions, the amount of PIT paid, the BT income[3], transfers and pensions not liable of PIT. All incomes can be computed at the individual and household level, allowing for different equivalence scales.

4 Building the tax base

As a first step the model has to put together the different types of incomes to build the tax base. TABEITA was initially built and developed following Dirimod95, using 1995 SHIW data (Mantovani, 1998)[4]. Analysis of data is performed starting from employment, pension, self-employment incomes, transfers, capital income and rents (respectively `crrld, crrpens, crrauto, crrtrasf, crrinterr, crrcat` stata programs). They all form the Italian PIT tax base and are aggregated by the `crrnetti` stata program.

Once all the components of the PIT tax base have been assembled, household relations are recovered identifying those who are required to present the tax form and those who are not, the family relations and the right to use tax allowances depending on family composition according to the tax code of the relevant year. This analysis is performed assuming a coincidence of household and family since the data do not allow one to disentangle the presence of more

[3]BT income is divided into five components: (a) employment, (b) self-employment, (c) pension, (d) rental and estate income, (e) capital, interests and participation.

[4]In particular TABEITA98 follows Dirimod95 quite strictly for reconstructing the BT income while the algorithm for obtaining AT income was developed ad hoc. However, so many were the changes made on Dirimod95, especially as for estate and rental incomes, grossing-up procedure and tax evasion estimation, that the TABEITA is quite different from Dirimod95, the latter bearing no responsibility for possible mistakes in the former.

than one family under the same dwelling.

Up to this point, income is only AT. The program then recovers the BT income excluding those components of AT income which are not liable to PIT (e.g. invalidity pensions), taking into account those tax allowances which do not depend on income, then applying an iterative procedure to recover numerically the BT income (`crdetded` and `crimpon` stata programs). For the Italian PIT code, the AT income of individual i can be derived as:

$$y_i^A = y_i^B - (\underbrace{y_i^B - yex_i}_{yc_i} - d_i)t_i + D_i + Dy_i \tag{2}$$

where t_i is the multiple-bracket tax schedule applied to taxable income; yc_i is the PIT gross income, which is equal to BT income minus PIT-exempt incomes, yex_i; d_i is a set of deductions to be subtracted from yc_i; D_i and Dy_i are a the set of tax allowances that do and do not depend on BT income, respectively, and that reduce the tax to be paid. The algorithm developed in TABEITA then recovers the BT income as:

$$y_i^B = y_i^A - D_i - Dy_i + (y_i^B - yex_i - d_i)t_i \tag{3}$$

Clearly, in the first step of (3) y_i^B and Dy_i on the RHS are unknown, then they are set to zero. From the second step onward, y_i^B on the RHS is replaced with y_i^B obtained in the previous iteration, as well as Dy_i is computed accordingly to y_i^B of the previous iteration. The process goes on until y_i^B does not change for two successive iterations for all individuals in the sample.

5 Grossing-up

As discussed in the introduction, MSMs are mainly used for forecasting and analyzing the impact of a change in the structure of the tax and benefit system on (a) the distribution of income and (b) national accounts.

For this purposes, it is necessary onto project the sample to country totals. This projection can be obtained using a simple proportion between the dimension of the sample and that of the national population, generally weighted using the sampling weights provided in the data set. More often, data sets come with weights to be used for national projections, which are obtained from a process of post-stratification of the sample to known population totals. Post-stratification is an issue that has been extensively analyzed in survey statistics (see for instance Sarndal et al. (1992)) and consists in calibrating some sub-samples (post-strata) of a data set to given totals. In the MSM literature post-stratification is more commonly referred to as grossing-up, since the problem consists in grossing the sample up to the population under study. The aim of the grossing-up procedure is to make the sample as close as possible to the true population, although it depends on the variable used for performing the grossing-up as well as on the procedure implemented.

The grossing-up procedure is basically aimed at adjusting the data set to reflect differential non-response between different groups in the sample. The grossing-up procedure consists in assigning to each unit in a sample of dimension N a weight p_j with $j = 1, ..., N$, such that some chosen statistics of interest calculated on the weighted sample coincide with the population statistics. The procedure is trivial if we want to reconcile the sample with the population using only one discrete statistic, s_k with $k = 1, ...K$, such as family types or income ranges. In this case, we compute the probability of having the

characteristic s_k in the sample, say $P(s_k)$, and make it equal to the probability of having the same characteristic in the population, say $p(s_k)$. If the dimension of the sample and of the population are N and n respectively, then the grossing-up weight is $p_j = np(s_k)/NP(s_k)$, i.e. the size of the cell with characteristic s_k in the population divided by the size of the cell with characteristic s_k in the sample. If more variables are considered for the grossing-up procedure it should be necessary to consider the interactions between the different variables, i.e. consider the joint distribution of the control variables considered. However, this conflicts with available information from external sources, that in general, do not report the joint distribution of population variables but only the totals for each variable. For instance, it is possible to know the total number of single-parent families and the total number of self-employed in the population but not how many single-parent family have self-employment income. Hence, the conditions imposed on the weights p_j are far less stringent than in the "full information" case we would have if the joint distribution were known, and in general there are many possible sets of weights p_j achieving the desired adjustment.

To choose among them Atkinson et al. (1988) suggest the requirement that given a data set of dimension N, with original sampling weights q_j, $j = 1, 2, ..., N$, the set of grossing-up weights p_j have the least deviation from original weights, q_j. The original weights could reflect the sampling procedure or be uniform. Both grossing-up and initial weights have to sum up to the population size: $\sum q_j = \sum p_j = n$. If original and sample weights sum up to the sample dimension, they first have to be multiplied by n/N. It is then common practice to impose the condition that the new weights minimize the distance from initial weights. Hollenbeck (1976) proposed to use as a measure of distance the half of the squared sum of the difference between final and

initial weights. However, in order to avoid negative weights, Atkinson et al. (1988) suggest minimizing a measure of distance derived from information theory (Theil, 1967; Cowell, 1980):

$$d(p,q) = \sum p_j log\left(\frac{p_j}{q_j}\right) \tag{4}$$

As for the optimal number of control totals to be included, no result is currently available. Although it is more common to face the problem of not having enough external sources than to have too many, Sutherland (1989, p. 15) warns on the risk of increasing the variance of weights since the larger the number of control totals becomes, the smaller the number of observations in each "cell' (i.e. with each combination of characteristics being controlled for).

Atkinson et al. (1988) applied their methodology[5] to TAXMOD, a MSM for the UK, and compared their results with what could be obtained with uniform weights, i.e. multiplying the sampling weights by n/N. The grossed-up results were significantly more plausible. The conclusion from their analysis is that the use of uniform weights can be seriously misleading.

The SHIW data set is post-stratified using the variables sex, age class, area and dimension of the town of residence (Banca d'Italia, 2006, p. 47). However, it is not clearly stated what methodology was used and, for instance, which age classes were considered.

Table 1 shows how much the weighted sample differs from population totals using the post-stratification weight provided in SHIW for year 2000. It can be seen that, using the SHIW weights, the differences between the post-stratified and actual figures are small (less than 1%) as far as sex and area of residence (North-West (NW), North-East (NE), Center (C) and South (S)) are consid-

[5]As control totals they used the variables (i) family composition, (ii) employment status, (iii) income range and (iv) housing tenure.

ered, but they become worryingly large for age groups (especially by area of residence) and schooling. Moreover, the post-stratified figures obtained using SHIW00 weights are rather different from those released by the tax administration. This is probably due to the fact that SHIW00 weights are calculated using the control totals provided by ISTAT. Since ISTAT control totals differ from those provided by the tax administration, there could be a problem with post-stratified simulations. For instance, the effects of an hypothetical tax policy that affected mainly the selfemployed would probably be overestimated as these groups are over-represented using the SHIW weights. All these issues are of particular relevance whenever an analysis of income by population subgroups is performed. For these reasons a set of alternative post-stratification weights were estimated using the same methodology as Atkinson et al. (1988) using control totals found in Ministero delle Finanze (2004) and ISTAT (2004). In particular the new post-stratified weights are consistent with ISTAT for the main socio-demographic characteristics, while they refer to tax administration data for the number and the geographical distribution of employed and self-employed. In this way the weighted SHIW00 can be considered consistent with the Italian population of taxpayers (see table 1) and can be used to simulate the redistributive effects of fiscal policy in that year. Summary statistics for the initial SHIW and grossed-up weights are reported in table 2. Similar grossing-up procedures are performed in each version of TABEITA.

In comparable Italian MSMs the issue of estimation of grossing-up weights alternative to those provided in the data set is often overlooked. Neither MASTRICT (Proto, 2000), nor Dirimod95 (Mantovani, 1998) and its updated version Mapp98 (Baldini, 2001), nor the Italian module in EUROMOD (Atella et al., 2001) address the problem and the weights provided in the SHIW data set are used instead.

Variable	Ext sources*	SHIW weight		Our Weight	
	Totals (a)	Totals (b)	Diff (b/a-1)	Totals (c)	Diff (c/a-1)
Total population	57844017	57828424	0.00%	57844127	0.00%
Males	27796000	28068065	1.00%	27796064	0.00%
Females	30048017	29760359	-1.00%	30048063	0.00%
Pop NW	15153050	15151408	0.00%	15153060	0.00%
Pop NE	10681233	10645864	-0.30%	10681278	0.00%
Pop C	11159583	11108163	-0.50%	11159577	0.00%
Pop S	20850151	20922989	0.30%	20850212	0.00%
Age<=19	11349415	11495187	1.30%	11349436	0.00%
19<Age<=65	35938667	35962509	0.10%	35938782	0.00%
Age>65	10555935	10370728	-1.80%	10555909	0.00%
Age<=19 NW	2562196	2837546	10.70%	2562209	0.00%
Age<=19 NE	1814818	1949195	7.40%	1814820	0.00%
Age<=19 C	1983300	2150925	8.50%	1983298	0.00%
Age<=19 S	4989101	4557521	-8.70%	4989109	0.00%
19<Age<=65 NW	9654836	9644351	-0.10%	9654847	0.00%
19<Age<=65 NE	6752727	6805086	0.80%	6752770	0.00%
19<Age<=65 C	6969449	7061154	1.30%	6969457	0.00%
19<Age<=65 S	12561655	12451918	-0.90%	12561708	0.00%
Age>65 NW	2936018	2669511	-9.10%	2936004	0.00%
Age>65 NE	2113688	1891583	-10.50%	2113688	0.00%
Age>65 C	2206834	1896084	-14.10%	2206822	0.00%
Age>65 S	3299395	3913550	18.60%	3299395	0.00%
Employed**	17723376	15563640	-12.20%	17723450	0.00%
Employed NW **	5278461	4600891	-12.80%	5278487	0.00%
Employed NE**	3843039	3408590	-11.30%	3843050	0.00%
Employed C**	3398935	3272614	-3.70%	3398947	0.00%
Employed S**	5202941	4281545	-17.70%	5202966	0.00%
Self-employed**	5837998	6471973	10.90%	5838009	0.00%
Self-employed NW**	1827274	1763441	-3.50%	1827278	0.00%
Self-employed NE**	1296121	1442827	11.30%	1296120	0.00%
Self-employed C**	1233728	1396112	13.20%	1233725	0.00%
Self-employed S**	1480875	1869593	26.20%	1480886	0.00%
Employed	15131000	15051997	-0.50%	17077358	12.90%
Employed NW	4616000	4464179	-3.30%	5091679	10.30%
Employed NE	3247000	3302600	1.70%	3712218	14.30%
Employed C	3050000	3213664	5.40%	3338018	9.40%
Employed S	4218000	4071554	-3.50%	4935443	17.00%
Self-employed	5949000	5867783	-1.40%	5216206	-12.30%
Self-employed NW	1678000	1598946	-4.70%	1614862	-3.80%
Self-employed NE	1367000	1297700	-5.10%	1146121	-16.20%
Self-employed C	1205000	1302585	8.10%	1147294	-4.80%
Self-employed S	1699000	1668552	-1.80%	1307929	-23.00%
Elementary schooling	19766000	19628313	-0.70%	19237834	-2.70%
Compulsory schooling	16556000	15624763	-5.60%	15666562	-5.40%
High school degree	14291000	15436445	8.00%	15732990	10.10%
Laurea	3546000	3799227	7.10%	3944788	11.20%
Agricolture	1120000	1310029	17.00%	1369843	22.30%
Industry	6767000	7051688	4.20%	7576260	12.00%
Services	13193000	12362652	-6.30%	13163325	-0.20%

Source: Our calculations on SHIW00

*External sources from ISTAT (2004) where not differently specified;

** External sources from the Tax Administration (Ministero delle Finanze, 2004)

Table 1: Dicrepancies between population totals and weighted sample. Year 2000.

	Obs.	Mean	Std. Dev.	Min	Max
SHIW	22268	2596.929	2442.227	303	18277
Our Weight	22268	2597.634	2477.708	248	26122

Table 2: Summary statistics for initial SHIW and final grossing-up weights.

6 Estimation of tax evasion and validation

A common finding from the SHIW data set is that, in some cases, income in the survey is on average higher than what declared to fiscal authorities and that the difference is larger for some incomes (e.g. self-employment) and smaller for others (e.g. employment). In other cases (for example, capital income), due to the presence of underreporting, income tends to be underestimated. Disregarding this fact would then imply a simulation of a different tax yield than that actually obtained, hence an incorrect forecasting of redistributive and revenue effects of different fiscal policies. Any MSM for Italy that uses SHIW data, needs to consider this discrepancy. The common practice is to assume that any positive difference comes from the fact that taxpayers are more honest with an interviewer that grants anonymity than with the fiscal authorities. The positive difference between the total amount of income grossed-up from individual incomes declared in the SHIW and the total amount of income declared to the fiscal authorities is therefore attributed to tax evasion or tax avoidance.[6] This approach has also been used in recent year to provide an estimate of tax evasion to be compared with other methods of tax evasion estimation (see, among others, Fiorio and D'Amuri (2006)). Alternative methodologies include the direct approach, the indirect approach or the latent variable approach[7] (for a review of results from different methods to estimate

[6] In this case, income variables are inflated using a constant percentage in order to meet the totals declared by the tax administration.

[7] The direct approach is based on tax audits or on census and labor force data and on analysis of expenditures. The assumptions are that the labor force participation obtained from population census and labor force surveys include unregistered employment, and that consumption is more truthfully declared than income, clearly an analogous assumption to what used here.

The indirect approach is based on the currency demand approach, which assumes that hidden transactions use cash. A demand for currency is than estimated with regression methodologies (for a review of the approach and updated results, see Schneider (2000)).

The latent variable approach considers the underground economy as a non observable variable and estimates the links with a set of determinants, including the production and

underground economy in Italy, see Zizza (2002)).

In Dirimod95 an estimate of evasion/avoidance is obtained making use of the analysis of 1991 tax forms, relative to 1990 incomes (Ministero delle Finanze, 1995). Tax evasion and tax avoidance is estimated in two stages. At first the model is run assuming zero evasion, data are grossed-up and total BT income is compared to BT income as found from aggregated tax forms. In the second stage increasing level of tax evasion and tax avoidance is estimated and the resulting income is compared with data from aggregate tax forms. In Dirimod95, however, tax form data refer to 1990 incomes and their updating to 1995 using a constant consumer price index (CPI) adds a bias in the estimation procedure. The module of EUROMOD for Italy deals with tax evasion and tax avoidance in a similar way, using results from MASTRICT, a MSM developed at ISTAT (Proto, 2000). In particular, in the EUROMOD module for Italy employment income tax evasion and tax avoidance is estimated to 0%, self-employment income to 50% (Atella et al., 2001).

The methodology to estimate tax evasion in TABEITA is similar to the one in Dirimod95. Constant percentages of tax evasion are calculated for different types of income, comparing total net incomes reported by aggregated tax form data with those calculated by the MSM model. Imputation of tax evasion is based on the tax form data from the same (or the nearest available) year.

Once the imputation of tax evasion and correction for underreporting is carried out, it is then possible to provide a first validation of the model's output. The validation procedure is carried out in each version of TABEITA, in Table 3 we show for brevity only the results obtained in year 2000. It can be seen that the results of the validation are quite reassuring about the reliability of TABEITA: discrepancies between values calculated by the MSM

the labor market activities. It employs latent variable econometric tools combined with factorial analysis (see, for instance Frey and Weck-Hanneman, 1984)

	Tabeita00 (a)	External Source (b)	Difference (a/b -1)
Gross inc.	555928	562309	-1%
Employment inc.	425870	426481	0%
Self-employment inc.	66334	67134	-1%
Rental/Estate inc.	27612	27696	0%
Capital inc.	36112	36539	-1%
Taxable inc.	532823	546670	-3%
PIT (Gross)	135060	136863	-1%
PIT (Net)	113693	108144	5%

Table 3: Validation of TABEITA00 Output, millions of euros. Our calculations using TABEITA00 and tax administration data.

and the official figures released by the tax authorities are between ±5%. This is true for aggregate and taxable income (i.e. aggregate income net of tax deductions) as well for gross PIT. The main discrepancy is found for net PIT, that is slightly overestimated by TABEITA (+5%).

7 Equivalence scales

Once the BT incomes have been recovered and all the consistency and reliability checks have been carried out, it is possible to exploit the MSM to perform an assessment of the redistributive effect of different PIT codes. Since the focus of the analysis is on household welfare, the use of an equivalence scale is necessary. Given the impossibility of obtaining a unique equivalence scale (see Cowell and Mercader-Prats (1997) and Blundell and Lewbel (1991)) the Italian Poverty Commission approach, which is derived from the Engel methodology, is employed by TABEITA. The elasticity of total consumption on family magnitude is estimated by a weighted regression where the dependent variable is the proportion of food expenditure on total expenditure (c_f) and, as independent variable, the log of total household expenditure (C) and the log of the number of the member of the household (N): $c_f = a + blnC + clnN + u$. The

elasticity estimate is consequently obtained as $\varepsilon = (-c/b)$ and the equivalent income of each member of household j can be estimated as: $y_h = x_h/N^{\varepsilon}$ where x_h is the household income and N is the number of household members.[8]

8 An application: estimation of the redistributive impact of the 2004 and 2005 Italian PIT reforms.

In 2004 and 2005 two subsequent PIT reforms were introduced. In 2004 an increase of the tax rates on low income brackets was accompanied by the introduction of a No Tax Area, decreasing in the level of gross reported income, and in replacement of work income tax allowances. In 2005 lower tax rates for high incomes were introduced, together with another No Tax Area (in substitution of family tax allowances) decreasing in the level of gross income and increasing in the number of non working relatives. The analysis of the overall distributive effects of these fiscal reforms is not straightforward: some modules of the reforms benefit low income individuals, while others are aimed at decreasing the tax burden on richer taxpayers. Moreover, the extent of the tax cuts varies depending on the composition and the size of the family and on the type of work income perceived. Finally, low income individuals benefit from the introduction of generous tax areas, but are also hit by the increase in tax rates on lower income brackets. Without a MSM the only possible way to try to understand the effects of the reforms is to refer to the study of "representative households", but this kind of analysis is obviously very limited in its scope.

[8]For a detailed discussion for the equivalence scale choice by the Poverty Commission, see De Santis (1998).

TABEITA can instead disentangle the overall redistributive impact of the tax reforms, simulating the change in taxes due on a sample representative of the population of taxpayers (see D'Amuri et al. (2004)). The results obtained with the MSM model can also be combined with the use of common indices of redistribution and with non-parametric density estimation technics. For the 2004 and 2005 tax cuts the redistributive effects were analyzed by means of the Reynolds-Smolensky index and its decomposition in incidence and Kakwani indices. To provide a counterfactual the same indices were computed using the AT distribution obtained simulating the 2001 PIT code on the same BT income distribution. As Table 4 shows, the 2004 and 2005 tax cuts lowered the overall redistributive effects of the Italian PIT: the Reynolds-Smolensky index decreased from 0.0371 (2001 PIT) to 0.0362 (2005 PIT). This result was due to a decrease in the incidence of the income tax, only partially offset by an increase in its progressivity. Figure 1 shows the non-parametric density estimation for average tax rates resulting form the application of the three different tax codes analyzed. The impact of the 2004 and 2005 tax cuts can be now clearly understood: the 2004 PIT reforms benefited only low and middle income taxpayers (with BT income lower than 35,000 euros), while in 2005 the tax burden decreases for middle and high income individuals (benefits start at 20,000 euros).

	2001 PIT	2004 PIT	2005 PIT
Reynolds - Smolensky	0.0371	0.0388	0.0362
Incidence	0.1928	0.1837	0.1728
Progressivity (Kakwani)	0.1555	0.1724	0.1735

Table 4: Redistributive effects of the Italian PIT, various years.

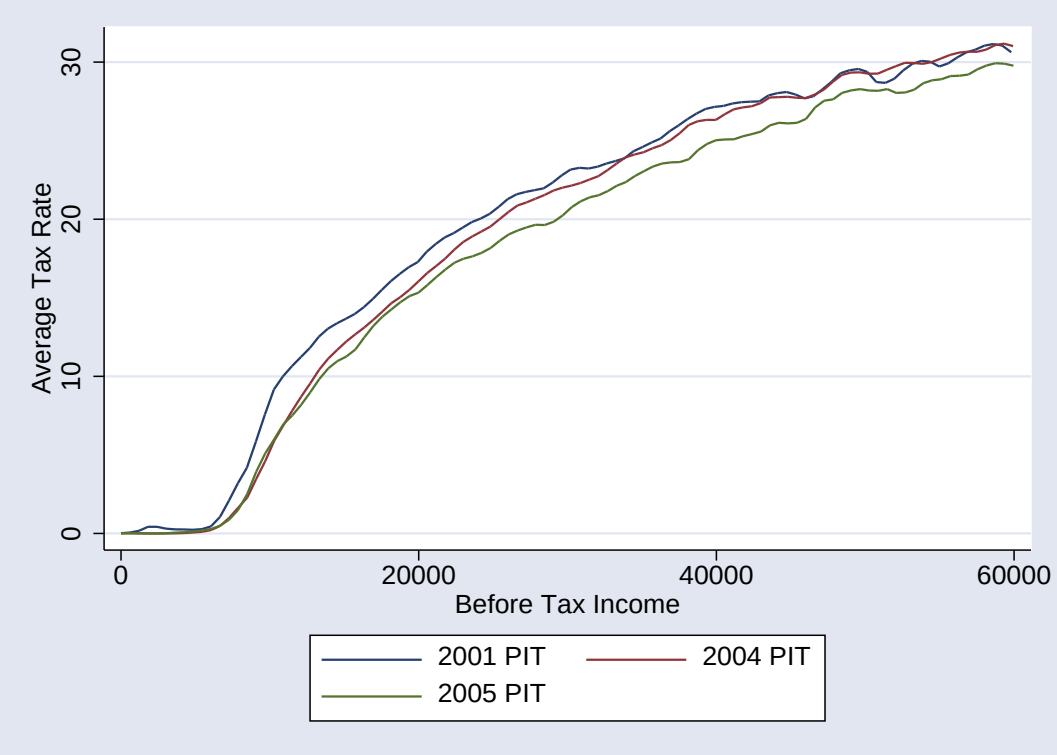

Figure 1: Estimated Average Tax Rates (%), Italian PIT, various years. Horizontal axis: values in Euros.

Appendix

A concise guide for TABEITA users

TABEITA is computed in STATA. The MSM runs on SHIW data provided about every two years by the Bank of Italy. The dataset has to be downloaded in the folder `C:/shiw/shiw04`[9] for 2004 data, while the STATA files constituting the MSM have to be saved in the following folder:

[9]For year 2002 use c:/shiw/shiw02 and so on.

`C:/mydocuments/projects/MSM/TABEITA04/stata/gross`.

The model is made by a series of STATA programs (see Table 5), each one can be run separately. The file `master.do` initializes the model and runs all its files subsequently. The first set of files creates a dataset with all the information on the individual AT income (see Table 5). `Crrevas` estimates different levels of income concealment for each type of income (employment, self-employment and estate income) and reconstructs the actual level of income reported to the tax authority. `Crdetded` calculates tax allowances, and finally `crimpon` reconstructs BT incomes by means of an iterative algorithm. After all the files have been run, a new dataset is created, containing all the relevant socioeconomic information, together with the new variables calculated by the MSM: gross income for every type of income, tax allowances, income tax actually paid. Once gross income has been calculated, simulation of the effects of changes in the tax code can be carried out. The model is actually equipped with programs that can simulate the Italian PIT from year 1998 to year 2005.

File	Description
master.do	Initializes the MSM and runs all the other programs subsequently.
crrld.do	Reconstructs net employment income. Consistency checks are carried out to control the accuracy of data aggregation and the extent of data imputation.
crrpens.do	Reconstructs net pension income. Consistency checks are carried out to control the accuracy of data aggregation and the extent of data imputation.
crrauto.do	Reconstructs net self-employment income. Consistency checks are carried out to control the accuracy of data aggregation and the extent of data imputation.
crrtrasf.do	Reconstructs income from welfare benefits, scolarships, family allowances etc. Consistency checks are carried out to control the accuracy of data aggregation and the extent of data imputation.
crrinterr.do	Reconstructs active interest received and passive interest borne by the individual. Consistency checks are carried out to control the accuracy of data aggregation and the extent of data imputation.
crrcat.do	Reconstructs real estate/rental income following the Italian Tax Code. Imputation of cadastral income is carried out following Baldini (2001), pp. 13-15. Consistency checks are carried out to control the accuracy of data aggregation and the extent of data imputation.
crrtot.do	Aggregates incomes by income type. Labels all main variables.
crrnetti.do	Aggregates incomes by type of taxable income.
crrevas.do	Imputes different levels of tax evasion by income type.
crdetded.do	Imputes tax allowances.
crimpon.do	Imputes BT income by means of an iterative algorithm.
irpef.do	Calculates allowances, taxable income, gross and net PIT following the tax code of the relevant year.

Table 5: TABEITA Stata programs description.

References

Atella, V., Coromaldi, M., and Mastrofrancesco, L. (2001). EUROMOD Country Report. Italy. *mimeo*.

Atkinson, A. B., Gomulka, J., and Sutherland, H. (1988). Grossing-up FES data for Tax-Benefit Models. Number 10 in STICERD Occasional Paper, pages 223–253. STICERD, London.

Baldini, M. (2001). Mapp98: un Modello di Analisi delle Politiche Pubbliche. *CAPP, Materiali di discussione, Modena*, (331).

Banca d'Italia (2006). *I bilanci delle famiglie italiane nell'anno 2004*. Supplementi al Bollettino Statistico. Note metodologiche e informazioni statistiche. Banca d'Italia, Roma.

Blundell, R. and Lewbel, A. (1991). The information content of equivalence scales. *Journal of Econometrics*, 50:49–68.

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

Brandolini, A. (1999). The Distribution of Personal Income in Post-War Italy: Source Description, Data Quality, and the Time Pattern of Income Inequality. *Giornale degli Economisti e Annali di Economia*, 58:183–239.

Cannari, L. and D'Alessio, G. (1992). Mancate interviste e distorsione degli stimatori. *Temi di discussione del Servizio Studi*, (172).

Cavalli, M. and Fiorio, C. V. (2006). Individual vs family taxation: an analysis using TABEITA04. *Econpubblica Working Paper Series*, 116.

Cowell, F. A. (1980). On the structure of additive inequality measures. *Review of Economic Studies*, 47:521–31.

Cowell, F. A. and Mercader-Prats, M. (1997). Equivalence of Scales and Inequality - Distributional Analysis Discussion Paper. Technical Report 27, STICERD, London School of Economics, London.

D'Amuri, F. (2004). *Equit verticale ed orizzontale dell'Irpef italiana. Analisi empirica con un modello di microsimulazione*. Tesi di Laurea. Bocconi University.

D'Amuri, F., Fiorio, C., and Zanardi, A. (2004). Il doppio passo della riforma aumenta i vantaggi. *Il Sole 24 Ore*. 3rd of December.

De Santis, G. (1998). Le misure della povertà in Italia: scale di equivalenza e aspetti demografici. In *Commissione di indagine sulla povertà e sull'emarginazione*. Presidenza del Consiglio dei Ministri, Dipartimento per gli affari sociali, Rome.

Duclos, J. (1995). Modelling the take-up of state support. *Journal of Public Economics*, 58:391–415.

Fiorio, C. and D'Amuri, F. (2006). Workers' tax evasion in Italy. *Giornale degli Economisti e Annali di Economia*, 64:241–264.

Fiorio, C. V. (2004). *Microsimulation and analysis of income distribution: an application to Italy*. PhD thesis, London School of Economics, Economics Department, London, UK.

Frey, B. S. and Weck-Hanneman, H. (1984). The hidden economy as an 'unobserved' variable. *European Economic Review*, 26:33–53.

Fry, V. and Stark, G. (1993). The Take-up of Means-Tested Benefits, 1984-90. Institute for Fiscal Studies, London.

Hollenbeck, K. (1976). An algorithm for adjusting n-dimensional tabular data to conform to general linear constraints. In *Proceedings of the American Statistical Association*, pages 402–405.

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

Ministero delle Finanze (1995). *Analisi del 740. Anno 1991 redditi 1990*. Ministero delle Finanze, Rome.

Proto, G. (2000). Il modello di microsimulazione MASTRICT: struttura e risultati. *mimeo*. ISTAT, Direzione Centrale Imprese e Istituzioni.

Pudney, S. (2001). The impact of measurement error in probit models of benefit take-up. *mimeo*, page University of Leicester: Working Paper.

Pudney, S., Hernandez, M., and Hancock, R. (2002). The welfare cost of means-testing: pensioner participation in income support. *mimeo*.

Pudney, S. and Sutherland, H. (1994). How reliable are microsimulation results? An analysis of the role of sampling error in a U.K. tax-benefit model. *Journal of Public Economics*, 53:327–365.

Sarndal, C. E., Swensson, B., and Wretman, J. (1992). *Model Assisted Survey Sampling*. Springer-Verlag, New York.

Schneider, F. (2000). The increase of the size of the shadow economy of 18 OECD countries: some preliminary explanations. *IFO Working papers*, (306).

Sutherland, H. (1989). Constructing a Tax-Benefit Model: What Advice Can One Give? In *Taxation, incentive and the distribution of income*, number 141. STICERD, London School of Economics.

Theil, H. (1967). *Economics and Information Theory*. North-Holland, Amsterdam.

Zizza, R. (2002). Metodologie di stima dell'economia sommersa: un'applicazione al caso italiano. *Temi di discussione del Servizio Studi*, (463). Banca d'Italia.