Income Tax Statistics Analysis: A Comparison of Microsimulation Versus Group Simulation

Heiko Müller[1] **and Caren Sureth**[2]
[1] Ruhr-University of Bochum, Faculty of Economics, Universitätsstr. 150, 44801 Bochum, Germany, and arqus, Quantitative Tax Research, www.arqus.info; email: Heiko.Mueller@ruhr-uni-bochum.de
[2] University of Paderborn, Faculty of Business Administration and Economics, Warburger Str. 100, 33098 Paderborn, Germany, and arqus, Quantitative Tax Research, www.arqus.info; email: csureth@notes.upb.de

ABSTRACT: Microsimulation based on income tax statistics may be useful in tax reform discussions. Unfortunately, access to appropriate data is still rather restricted and expensive for ad-hoc analyses, and individual data is often even not available at all. In this paper we take Germany and its data situation as a proxy for many countries' restrictions in terms of tax data availability. Analyzing how much reliability and robustness of results we lose if we employ group simulation instead of microsimulation, we compare both methods. Investigating tax scale effects by the group model leads to very good results. Determining the financial effects of modified tax bases, the deviation from the microsimulation results increases, especially if tax base cuts vary between taxpayers. In addition, we take account of the class of taxpayers with a negative taxable income. Neglecting this class we identify a systematic underestimation of the financial consequences of a modified tax base with the group model assuming a progressive tax scale. If the group simulation data is not arranged according to the taxable income, but rather according to the total amount of income, we also find a tendency towards higher deviations from the microsimulation results. Quantifying the tax revenue effects of alternative tax settings the group simulation model represents a good compromise between the desire to capture the complex reality and the achievable accuracy when facing limited resources and data. Furthermore, for those cases in which group simulation is the appropriate tool, we provide a very simple method to interpolate a suitable income distribution and thereby the tax distribution within the classes. This interpolation makes future estimates of tax revenues a lot easier. We conclude that, although microsimulation in general is the superior approach, a group simulation model remains of interest, especially for analyses of rather old data and cross-country analyses, when sufficiently detailed data for micro analyses is missing.

Keywords: microsimulation; group simulation; tax revenue; personal income tax; tax statistics

I. INTRODUCTION

Microsimulation models of income tax systems are usually employed to analyze the fiscal and distributive issues of taxation. These are important fields of research. The results may be useful in tax reform, budget and income distribution discussions and therefore may contribute substantially to solving these three major economic questions. As long as complete microdatasets are available microsimulation is the preferable tool. However, access to appropriate data in a number of countries is still rather restricted or expensive for *ad hoc* analyses. In addition, in the case of analyses based on data from previous assessment periods individual data are often not available at all. Even in the industrialized countries data collected earlier than 10 to 15 years ago is usually not microdata but grouped data. Consequently analyses of times series often have to fall back on group data. This is true even for industrialized countries where microsimulation models have become a widespread tool for the analysis of newly collected data. Hence, group simulation models often have to be applied for specific countries, for cross-country analyses and for long-term time series analyses. Against this backdrop it is important to find out how robust the results from group simulation are and thus how big is the error arising from the more aggregate group model in comparison to a microsimulation model. Conversely, given the expense and effort involved in setting up a microsimulation model, it is also important to consider under what circumstances sample-based microsimulation that uses incomplete microdatasets remains superior to group simulation.

After the amendment of the German Act on Fiscal Statistics in 1996 it was for the first time possible to consolidate the individual data records from the local statistical offices centrally and to use them for auxiliary and special analyses (cf. Zwick, 2001:640, see further Dell, 2007). Now the data can be prepared more flexibly and used for microsimulations for research and policy purposes. However, because of the generally limited access to microdata or for reasons of economy it is sometimes recommendable for several types of analyses of tax revenue effects to refer instead to classified data from income tax statistics.

In the following investigation we take Germany and its data situation as a proxy for many countries' restrictions in tax data availability. This analysis enables us to draw some general conclusions about how to deal with these limitations in future research in countries with a highly developed tax administration and tax statistics but insufficiently detailed and published tax data.

A vast body of literature examines the impact of income taxation on income distribution and tax revenues referring to different sources of data using either micro or group models. For an overview see, for example, Atkinson and Bourguignon (2000)

and Morrisson (2000).

Based on the seminal work of Orcutt (1957) the potential of microsimulation as a new analytical tool emerged. Orcutt, Merz and Quinke (1986) and Citro and Hanushek (1991) provide contributions of various authors and describe the opportunities and limitations of research based on microsimulation models for policy support purposes. With specific relevance to this paper, Cowell (1984) and Zandvakili (1994) examine microdata from household surveys to identify redistributive effects of taxation, whilst Merz (2000) employs sampled microdata from the German income tax statistics to analyze the redistributional impact of the German tax system. Bork and Petersen (2000), Wagenhals (2001) and Haan and Steiner (2005) similarly employ microsimulation to analyze German tax reform effects. A more detailed overview of the recent literature on microsimulation models relying on German data is provided by Wagenhals (2004).

Further research applying microsimulation tax-benefit models, based on microdata from several countries, provides a deep insight into the tax effects of varying taxation systems. Sutherland (1995) gives an overview of static microsimulation models in five European countries and prepares the field for a European model. Callan and Sutherland (1997) explore the prospects and limitations of such models referring to a case study. They point out that the level of detail inherent in a micro model based on microdata allows researchers to adjust simulations for transnational approaches. But still, this challenge is very demanding as differences in data availability, quality and definitions may have an impact on the results of each country. For example, Atkinson (2007) stresses that in specific cases tax data may be superior to data from household surveys employing UK tax data, whilst Pudney and Sutherland (1994) discuss the reliability of microsimulation results and show that sampling error in micro models often can be very significant. On the other hand, Zandvakili (1994) points out that microdata is usually superior to aggregated data with comparable variable definition.

In contrast to the microsimulation literature, Kakwani (1977) focuses on the problem of measuring progressivity in taxation and public expenditure and conducts an inter-country comparison using group data from the official income tax statistics. Kraus (1981) employs such data as well to investigate income inequality. Loizides (1988) also uses group data from the official Greek tax statistics to measure progressivity effects. Differences between twelve OECD countries are identified by Wagstaff, van Doerslaer, van der Burg et al. (1999) and Wagstaff and van Doorslaer (2001) using household survey and grouped OECD data. Piketty (2003) highlights French tax data deficits and estimates income inequality in France on the basis of tax statistics. Piketty and Saez (2003, 2007) and Saez and Veall (2007) look at US and Canadian grouped tax data. Dell (2007) uses group data from the German tax return statistics,

identifies several breaks in data over time, and stresses certain limits of recent data on tax bases and taxes paid. All of them investigate the tax impact on distribution, especially on top incomes over the twentieth century.

Whereas several papers point out that using group data limits the reliability of their studies in general (cf. Kakwani, 1977:75, Orcutt, 1982, Caldwell, 1985, McClung, 1986, Wagstaff and van Doorslaer, 2001:313), there is no analysis about the extent of inaccuracy arising from data deficiencies. One aim of our paper is to partly fill this void.

As the instrument of microsimulation cannot always be applied, due to either lack of data or resource, it is also of interest to consider how much reliability and robustness of results we lose if we use group simulation instead. A second aim of our paper, therefore, is to compare the outcomes of both methods. We apply microsimulation to microdata and group simulation to classified data, drawn from the same underlying dataset. The results allow us to draw conclusions about the opportunities and limitations of group simulation compared to those of microsimulation models. Specifically, we are able to show under which circumstances microsimulation is undoubtedly the superior approach and when group models provide reasonable estimates. Furthermore, we are able to show that in those cases in which group simulation is an appropriate tool, a very simple method to interpolate a suitable income distribution and thereby the tax distribution within the classes can be applied. This result makes future estimates of tax revenues a lot easier.

The remainder of this paper begins with an introduction to the tax statistics of the German Federal Statistical Office in section II. In section III we describe the main characteristics, advantages and limitations of micro and group simulation models. We present our model in section IV and the simulation results in section V. On this basis we summarize and draw final conclusions on the applicability, reliability and robustness of results obtained from the alternative methods in section VI.

II. TAX STATISTICS OF THE FEDERAL STATISTICAL OFFICE

One aim of this paper is to compare microsimulation models with group models and identify settings for which one or the other is preferable. In contrast to group models microsimulation is highly demanding of both data and human resource inputs. In these circumstances, an alternative modelling tool that requires less detailed data and human resource input might well be more appropriate. Setting resource inputs to one side, the problem of data (non-)availability can be illustrated if we take a look at the example of the German Federal Income Tax Statistics.

Income statistics are secondary statistics, i.e. the tax authorities provide summary tax statistics

based upon data collected during the tax assessment procedure. These data are not collected through questionnaires but extracted from personal tax assessments recorded by the fiscal administration for statistical reasons. The income tax statistics, however, are only assembled every three years by the German Federal Statistical Office, with a time-lag of at least of four or five years.

A multitude of data from wage tax cards, tax returns and from official tax assessment notes are documented in the tax statistics. Married couples that are jointly assessed are regarded legally as one tax payer (c.f. Wagstaff and van Doorslaer, 2001:307 on the problems of a tax unit referring to an individual or a couple). The 1995 tax statistics contain approximately 30 million data records covering 38 million persons, with around 400 attributes per record (Zwick, 2001:641). Besides technical and socioeconomic information these attributes include the data necessary to determine the individual tax base and people's personal tax liabilities.

The German Federal Statistical Office publishes part of these data in tables that provide grouped information only. In these tables, group-specific information is given for sets of taxpayers within intervals of a given income definition, for example, for classes of "total income" or classes of "taxable income". Researchers do not have access to the complete microdatasets. For the years before 1992 microdata is not available at all for research purposes. Hence, analyses on earlier years have to fall back on grouped data. In contrast, for the years 1998 and 2001 the German Federal Statistical Office has provided researchers with access to scientific usefiles. These files contain a stratified sample of the complete microdata base compiled for microsimulation purposes.

In line with the official calculation procedure for tax assessment all adjustments to income declared taxable, such as allowances for special expenses and expenses for extraordinary financial burdens, are considered as legitimate. Beginning with the "income from different sources of taxable income" these adjustments are conducted and finally produce the tax base, i.e. the "taxable income". In addition to the tax base the tax liability is documented in the income tax statistics. Applying the tax scale to the taxable income leads to the "tax scale income tax". Then, tax credits, tax pre-payments (for example, by wage tax and source taxes), tax refunds and so on have to be taken into account to arrive at the assessed tax liability.

The tables published by the German Federal Statistical Office distinguish between classes of "total income" and classes of "taxable income". The total amount of income is a kind of preliminary tax base, i.e. a tax base before special individual expenses and expenses for extraordinary financial burdens. The tables contain the aggregated value of the underlying attributes from the taxpayers'

microdata, for all tax payers or for certain selected groups of taxpayers, such as those subject to the basic or splitting tax scale. Whereas the base income tax scale is applied to individual taxpayers, married couples are subject to the splitting tax scale. To determine the income tax of a couple, the incomes of both spouses is summed and then halved. This halved income is subject to the basic income tax scale. The resulting income tax has to be doubled to calculate the couple's income tax. This procedure is called 'applying the splitting tax scale'.

Thus the published tables provide group-specific information about the tax base and the assessed tax. The tables used for group simulation only provide mean values for each attribute and class. When the underlying microdata have not been released, these grouped data have to be used instead. Overall a substantial information loss arises from aggregating data in each tax class in comparison to the corresponding individual tax microdata. The remainder of this paper investigates whether this information loss leads to high or rather negligible simulation differences.

III. MICRO VS. GROUP SIMULATION

Referring to the most important distinctive feature - the degree of aggregation of the applied data - in economics and the social sciences - we find three basic types of simulation model:

- Models that are essentially based on the aggregates from the national accounting system, like macroeconomic models and general equilibrium models (high aggregation level),
- Group models that refer to selected attributes of homogeneous groups of economic units (medium aggregation level), and
- Microanalytic models that focus on individual micro units (strong disaggregation).

Macroeconomic models and equilibrium models are not generally suitable for analyzing income tax revenue. In general equilibrium models a normally complex formula including a macroeconomic growth rate and other macro parameters are used to estimate the effects of monetary and fiscal policy on prices or employment or other macroeconomic variables. If such models are employed to estimate tax revenue effects relatively high prediction errors occur in comparison to more detailed approaches (group or micro models), due to the higher degree of aggregation. Attributes of the households, taxpayers and structural factors are insufficiently considered both in the model and in the results.

In comparison, the more intensively disaggregated group models and microanalytic models offer structural advantages. Generally, group models have a relatively simple and transparent structure compared with the microanalytic models. This facilitates their implementation and modification and makes them a flexible and low cost instrument for investigating revenue effects. This advantage

has to be offset against the previously mentioned information loss caused by using data aggregated with respect to a specific attribute. Hence the field of application of group models is restricted by the underlying aggregation pattern. If microdata are not available, the ensuing analytical limitations of group models have to be accepted. The question is whether the adverse impact of these limitations, in terms of analytical outcomes, is acceptable.

In contrast, if suitable data are available the higher degree of disaggregation that can be achieved using microanalytic simulation models is superficially desirable and necessary, whether analyzing the distributive effects of various tax and transfer systems or undertaking behavioural simulations. Microeconomic models take explicit account of taxpayers' individual attributes and hence allow us to determine the tax base and tax liability more precisely. It is therefore theoretically possible to make a more accurate and differentiated assessment of the revenue effects of, for example, a tax reform.

In a (pure) microanalytic simulation each individual micro unit with its attributes is referred to directly. This can be realised on the basis of individual cases, a sample or the parent population. The advantage of comprehensive and detailed structural information can only be exploited if an appropriate multiplicity of attributes of the micro units is available in the database. In order to achieve a simulation as close to reality as possible interdependencies of tax reform and individual behaviour have to be taken into account. Thus, we have to refer to the relevant elasticities, utility functions and so on in the model on either an empirical or theoretical basis. This increases the complexity of the model as well as the number of attributes.

Even if the microanalytic models are theoretically superior to the group models, the required specification and format of the data and the necessity to update it often limit or even prevent the application of microsimulations, particularly when dealing with long-term time series and cross-country analyses. In particular for *ad hoc* analyses or analyses of earlier tax periods we may have to fall back to the published aggregated data as no other detailed data is available. In these cases only group simulation models can be employed. In any case, micro models are often *de facto* group models, as data limitations sometimes necessitate the assumption that all individuals in the same group share the same attribute or distribution.[1] As a result scenarios can be identified for which group model results hardly differ from those of microdata analyses.

A disadvantage of group tax simulations is that they tend to lead to tax revenues that are too small. This is because progressive income taxation is usually not simulated correctly, a result of referring to aggregate income per income class and aggregate income tax per class instead of exact individual

income. In the case of microsimulation an empirical income distribution is inherent in the underlying microdatasets. In contrast, for group simulation purposes an empirical frequency distribution has to be formally estimated from the available aggregated data by applying specific distribution functions. Under these circumstances group simulation potentially becomes an attractive and powerful instrument and alternative to microsimulation models. This estimation of unknown empirical distributions can be achieved in principle by two methodological approaches:
- Applying analytic distribution functions whose parameters are derived from empirical material by approximation, or
- Applying interpolation functions.

In comparison to micro models one major drawback of group simulation models relying on an analytical distribution function is that the mathematical approximation of the analytic distribution functions to the unknown empirical distribution is very time-consuming and complex. Furthermore, there are often substantial deviations, in particular in the upper and lower income classes. It should also be noted that the advantage of using an analytic distribution function is often limited by the lack of a usable economic interpretation of the function parameters. If no acceptable mathematical approximation can be achieved we have to abstain from a theoretical approach to empirical income distribution and conduct an interpolation instead. In the next section of the paper we describe the construction of a group tax simulation model and, as part of this description, put forward one possible approach to approximating the empirical income distribution.

IV. THE MODEL

In the following, we introduce a discrete income tax simulation model based on classified data from German Fiscal Statistics.[2] The aim of this group model is to identify the revenue effects of alternative tax rules or systems, particularly the fiscal consequences of specific tax regulations, rapidly and flexibly. The group model is based upon available aggregate data from the income tax statistics. After presenting the group model we compare the results of micro and group simulation calculations in order to assess the relative accuracy of the group and microsimulation models and hence find out under which circumstances group simulation is or is not an appropriate approach, and in particular under what circumstances microsimulation models cannot be substituted by group models in an acceptable way.

IV.1 Discrete income distribution
Tax revenue analyses can normally be conducted without an analytic income distribution. Analytical theory-based income distributions only approximate real world distributions. Widely used analytic approximations include the log-normal (e.g., Berglas, 1971:534) and Pareto distributions (e.g., Piketty and Saez, 2003:6; Saez and Veall, 2007:

230). It is preferable, we argue, to deduce the income distribution directly from the available data. As data on the number of tax payers for each specific unit of taxable income *TI* is unavailable we need to interpolate. We derive the results presented in the following by applying a group simulation model and determining the distribution of income by means of a linear interpolation of the group simulation. An arithmetical series (i.e. a discrete function), rather than a continuous function, is chosen to approximate the distribution of income. Generating discrete income distribution functions is appropriate for tax revenue analysis since the domain of the income tax scale function contains only natural numbers and thus discrete arguments.[3] During the interpolation, the aggregate taxable income of all taxpayers in each tax class is also considered.

The discrete model presented for simulating personal income taxation based upon grouped data ensures that in each class aggregated taxable incomes and numbers of taxpayers are identical to the original microdata totals. Therefore, a degree of precision in disaggregation can be achieved that leads in each class to a 100% correct agreement between the aggregated taxable incomes and the amounts indicated in the tax statistics.

The absolute frequency of taxpayers with a specific taxable income *TI* is $h_{(TI)}$ and yields from the closed income interval *i* with the interval bounds $[a_i, b_i]$, with $a_{i+1}=b_i+1$ of the discrete density function of the taxpayers:

$$(1.1) \qquad h_i = \sum_{TI=a_i}^{b_i} h_{(TI)}.$$

Only the highest income interval has an open upper bound with $b_n=\infty$. This set of numbers (eq. 1.1) is a unique transformation of a set of natural numbers (taxable income) on a set of integers (absolute frequency of the taxpayers).

The sum of the taxable income of the taxpayers in the interval *i* is TI_i and can be determined as follows from the density function:

$$(1.2) \qquad TI_i = \sum_{TI=a_i}^{b_i} h_{(TI)}TI.$$

Applying the income tax scale to the tax base *TI*, neglecting preliminary special tax scale regulations, we receive income tax $t_{(TI)}$ [4]. The sum of the determined income tax of all taxpayers of the interval *i* is T_i and is given by:

$$(1.3) \qquad T_i = \sum_{TI=a_i}^{b_i} h_{(TI)}t_{(TI)}.$$

IV.2 Taxable income class
As already described, the published tables from the income tax statistics – separated into taxpayers underlying the basic scale and taxpayers underlying the splitting tax scale – include aggregate data for a

variety of tax relevant facts. An example is provided in Table 1. For each band of taxable income the number of taxpayers and relevant sum in DM is displayed.

For the purposes of tax revenue analysis it is appropriate to run a group simulation using data grouped with respect to classes of taxable income, since the range of values for the tax base of the taxpayers in each class is explicitly given and, thus, the interpolation of the distribution of the taxpayers is limited to this interval.

We use information relating to the number of taxpayers with a taxable income, the sum of the taxable income of these taxpayers and the sum of assessed income tax from the income tax statistics. This database can formally be described for taxpayers subject to the basic or splitting tax scale as follows:

Given are classes of "taxable income" *TI* for *i*=1 to *n* classes with the class limits $[a_i, b_i]$, where $a_1=-\infty$, $b_1=0$, $a_2=1$ and $b_n=\infty$. For every class *i* we know:

- the class frequency h_i (number of taxpayers of the class *i* for whom a taxable income has been assessed),
- the sum of the taxable income TI_i of the taxpayers of class *i*, and
- the sum of the assessed income tax AT_i of the taxpayers of the class *i*.

The assessed income tax AT_i of all taxpayers results from the application of all relevant tax rate regulations, tax reductions and tax base additions without imputable taxes.

Unfortunately, the income tax statistics do not include the "tax scale income tax" but the sum of the "assessed income tax" of each class. In contrast to the "assessed income tax" the "tax scale income tax" results from the assessment process at a stage before special regulations, tax reductions and tax base additions are considered. Furthermore, the absolute frequency of the taxpayers with a specific taxable income, $h_{(TI)}$, the sum of these taxable incomes, $h_{(TI)}TI$, as well as the corresponding income tax from $h_{(TI)}, t_{(TI)}$, cannot be found in the aggregate data of the income tax statistics. Only the average taxable income of each class,

$$\overline{TI_i} = \frac{TI_i}{h_i},$$

can be determined by dividing the sum of the taxable incomes and the number of taxpayers of the class. Further information that may be helpful to analyze the distribution of the taxpayers within the class is not available. Since the total assessed tax, *T*, is the result of assessment after considering all individual relevant tax regulations no additional information about the distribution of the taxpayers can be gained by referring to sums of assessed income tax in the respective income classes (AT_i) published in the income tax statistics.[5] Even if we

Table 1 Example of grouped data provided by the German Statistical Office, positive and negative income from different sources and assessed income tax (1995 income distribution, basic tax scale)

taxable income *from.. to under …DM*	negative income from different sources		positive income from different sources		Assessed income tax		Amount assessed using 1990 tax scale
	number of taxpayers	*sum in DM '000*	*number of taxpayers*	*sum in DM '000*	*number of taxpayers*	*sum in DM '000*	*sum in DM '000*
Under 1	266,105	-12,415,250	621,453	9,126,448	8,061	890	-
1 - 5,670	70,316	-710,317	1,722,122	12,318,096	166,060	54,234	-
5,670 - 8,154	29,302	-315,997	750,329	9,407,202	375,611	135,333	180,739
8,154 - 12,096	42,971	-495,175	1,133,053	18,737,618	633,921	529,002	965,929
12,096 - 12,366	2,634	-32,123	65,535	1,272,609	64,297	59,071	83,707
12,366 - 13,068	6,541	-72,121	159,406	3,180,959	155,468	171,071	219,226
13,068 - 18,036	48,221	-574,758	943,652	22,260,815	933,305	1,795,389	1,837,818
18,036 - 25,002	77,933	-939,391	1,217,176	36,818,996	1,215,761	4,031,413	4,018,826
25,002 - 30,023	61,778	-738,561	929,610	33,453,121	929,514	4,411,434	4,408,774
30,023 - 40,013	132,580	-1,389,631	2,146,192	91,273,023	2,146,147	14,380,689	14,380,662
40,013 - 50,004	113,496	-1,334,815	1,659,698	86,684,169	1,659,673	15,650,731	15,660,133
50,004 - 55,728	53,055	-612,021	599,575	36,517,684	599,574	7,160,721	7,173,713
55,728 - 58,644	21,470	-264,002	231,363	15,215,855	231,360	3,100,941	3,108,599
58,644 - 60,048	9,722	-135,733	98,916	6,747,897	98,916	1,399,886	1,403,312
60,048 - 66,366	38,500	-507,366	352,999	25,532,735	352,999	5,449,135	5,465,481
66,366 - 70,038	18,540	-282,451	152,869	11,943,497	152,866	2,644,972	2,652,440
70,038 - 75,006	21,053	-316,026	158,898	13,170,377	158,898	3,001,504	3,013,301
75,006 - 100,008	61,821	-1,177,064	365,395	35,597,184	365,365	8,730,816	8,784,647
100,008 - 120,042	20,876	-575,376	94,303	11,909,994	94,301	3,263,494	3,299,181
120,042 - 240,084	34,145	-1,370,692	107,606	19,967,383	107,577	6,268,572	6,458,174
240,084 - 480,168	9,642	-669,580	23,523	8,923,469	23,518	3,294,257	3,488,455
480,168 -1,000,026	3,404	-402,661	7,793	5,920,376	7,787	2,337,854	2,549,395
1,000,026 or more	2,059	-687,475	4,302	15,232,935	4,299	6,096,409	7,226,559
total	1,146,162	-26,018,587	13,545,766	531,212,441	10,485,277	93,967,818	96,379,068

Source: German Statistical Office, Wiesbaden.

assume identical tax bases for every taxpayer of an income class different income tax assessments may arise, as specific tax regulations may lead to different reductions and additions. A strict functional relation between the assessed income tax and the assessed tax base "taxable income" cannot be assumed.

In contrast to microsimulation, we hence have to be aware of the fact that during a simulation based on classified data the above mentioned problem for progressive income tax scale will occur. If we determine the income tax revenues referring to average taxable income per income class by multiplying the income tax on the average taxable income, $t_{(\overline{\pi}_i)}$, of the class with the number of taxpayers of the class, h_i, the deduced tax revenue will generally be too low. This is due to the fact that within the segment of the progressive rise of the income tax rate, the income tax on the average assessed tax base may not map the effect of the progressive structure precisely. Furthermore, the effects of a transition between two tax scale zones of the tax schedule cannot be reproduced within a class because the average taxable income of the

class can lie only in one zone.[6] This affects particularly the simulation of the revenues from reformed tax bases and reformed tax schedules with different tax scale zones.

In the following, in order to reduce these inaccuracies when determining income tax revenues by means of a group simulation based on classified data, we develop a discrete model for the taxpayer distribution within a class by applying linear interpolation (cf. Wagstaff and van Doorslaer, 2001:307; Atkinson, 2007:91-92; Saez and Veall, 2007:230). The linear interpolation requires the description of m elements between two numbers z_1 and z_2 with the difference $z_2 - z_1 = d$ in such *a way* that a finite arithmetic series of numbers emerges whose first element is z_1 and whose *(m+2)*th element is z_2. If d denotes the difference of the wanted arithmetical series of numbers, then

$$z_2 = z_1 + (m+1)\overline{d} = z_1 + d, \text{ i.e. } \overline{d} = \frac{d}{(m+1)}.$$

In a first step we assume that the taxpayers in the closed interval i (class) with the interval bounds

[a_i, b_i] are equally distributed. In this case the average taxable income of a class is identical to the mid-point of class:

$$\overline{TI_i} = \frac{(a_i + b_i)}{2} .$$

The sum of the taxable income of all taxpayers of a class is given according to eq. (1.2) by the product of the average taxable income and the number of taxpayers of this class:

$$TI_i = h_i \overline{TI_i} .$$

The aggregated income tax of the class can easily be determined by eq. (1.3) since the absolute frequency of the taxpayers for every taxable income within the interval is identical and can be described by

$$h_{(a_i)} = h_{(a_{i+1})} = ... = h_{(b_i \ 1)} = h_{(b_i)} = \frac{h_i}{(b_i - a_i + 1)} .$$

However, the average taxable income of a class is usually not equal to the mid-point of a class, meaning that the distribution of taxpayers within the class is obviously not uniform. In such a case, an assumption about the distribution of the taxpayers within the class is necessary.

Starting with the uniform distribution a discrete function (arithmetical sequence of numbers) that is strictly monotonously increasing or falling has to be assumed for the distribution of the taxpayers in the class. This function is conditioned on the position of the average taxable income in the class in relation to the mid-point of the class. We presume that the number of taxpayers in the mid-point of the class is equal to the quotient of the total number of the taxpayers of this class and the class breadth, i.e.

$$h_{\left(\frac{a_i + b_i}{2}\right)} = \frac{h_i}{(b_i - a_i + 1)} .$$

In this way, the problem is reduced to redistributing a certain number of taxpayers between the lower and upper class halves so that the sum of the income of the class corresponds to the empirical value. This redistribution is standardized such that the number of taxpayers at the beginning and end of the class differ exactly by two taxpayers, i.e.

$$\left| h_{(a_i)} - h_{(b_i)} \right| = 2 .$$

Thus, the difference between the number of taxpayers in the mid-point of the class and the number of taxpayers at the class beginning or the class end is exactly:

$$\left| h_{(a_i)} - h_{\left(\frac{a_i + b_i}{2}\right)} \right| = \left| h_{(b_i)} - h_{\left(\frac{a_i + b_i}{2}\right)} \right| = 1 ;$$

in other words, one taxpayer.[7]

Within the class the number of taxpayers rises and falls with $2/(b_i - a_i)$ whenever the underlying taxable income *TI* is amended by one DM.[8] This standardization ensures the required strict monotony. The degree of redistribution within a class u_i, can now be determined by referring to the empirical taxable income of the class:

(1.4)
$$u_i = \frac{TI_i - \frac{a_i + b_i}{2} h_i}{\sum_{TI=a_i}^{b_i} \frac{TI - a_i - \frac{b_i - a_i}{2}}{\frac{b_i - a_i}{2}} TI} .$$

The number of taxpayers with a specific taxable income under the given set of assumptions is:

(1.5)
$$h_{(TI)} = \frac{h_i}{b_i - a_i + 1} + \frac{TI - a_i - \frac{b_i - a_i}{2}}{\frac{b_i - a_i}{2}} u_i.$$

Considering u_i and h_i the number of taxpayers, $h_{(TI)}$, and thereby $h_{(TI)}$ *TI* and $h_{(TI)}$ $t_{(TI)}$ can be estimated for every taxable income. Inserting the frequencies of the taxpayers from eqn. (1.5) into the eqns. (1.2) and (1.3) we find for every class *i* that the sum of taxable income TI_i equals exactly the empirical value from the income tax statistics. This is true since $h_{(TI)}$ is determined via TI_i. Furthermore, the total income tax T_i of this class can be estimated.

Proceeding like this when determining the aggregate income tax of a class we succeed in reducing the systematic underestimation in group models fundamentally. If the aggregate tax of a class is determined by multiplying the income tax on the average taxable income of the class with the number of taxpayers of the class under a progressive tax, we receive the minimum level of the possible total tax of the class. If we instead employ a strictly monotonous discrete function that is defined on the basis of the empirically determined number of taxpayers and the sum of the taxable income of the class, then the total tax of a class varies between the theoretical minimum and maximum possible total tax of this class.

V. COMPARING TAX REVENUES EFFECTS OF MICROSIMULATION AND GROUP SIMULATION MODELS

V.1 Tax scale simulation based on taxable income

This type of group simulation allows us to obtain quite exact results involving relatively low effort, particularly when simulating different tax scales. The quality of this simulation approach can be emphasized in the following by comparing the results of a microsimulation, carried out by the

German Statistical Office, with those of the discrete group simulation model introduced here. The simulations of the German Statistical Office consulted for comparison purposes were carried out on the base of individual datasets from a 10% sample of the 1995 income tax statistics. The 10% sample is a formally anonymizised sample taken from the entirety of the recorded income tax assessments of the 1995 assessment period in the income tax statistics. This sample is a stratified random sample provided by the German Statistical Office.

In the following, the simulation of tax patterns is stylized, i.e. aligned with the main characteristics of the tax code. Thus, specific regulations, such as German tax relief for commercial earnings applicable only in 1995, have been neglected. The initial values of the sample and the results of the sample from the simulation were extrapolated to the parent population by the German Statistical Office. On basis of the aggregated data of the extrapolated initial values for the number of taxpayers and the aggregated taxable incomes of the classes, we run simulations using the discrete group model. Since the German Statistical Office defines the lowest income class as having no lower and the upper as having no upper limit, these class borders for group simulation purposes are heuristically determined. Therefore, assuming a uniform distribution, the average taxable income of the class is equated with the mid-point of the class:

$$\overline{TI_i} = \frac{a_i + b_i}{2} \ .$$

Thus, the upper limit of the interval is equivalent to twice the mid-point of the class, i.e.

$$\overline{TI_i} \times 2 = b_i \ .$$

This also applies to the lower class limit of the first class, i.e.

$$\overline{TI_i} \times 2 = a_i \ ,$$

because this class contains all taxpayers with a taxable income of less than one DM and therefore, the taxable income may even be negative in this class.[9]

The results presented in Table 2 show that the differences between the results from using our group simulation model and the results from the simulation conducted by the German Statistical Office based upon theirs sample microdata, both applying the basic tax rate and the 1990 and 1996 income tax scales, are very small. This result is robust even if we analyze the splitting tax scale instead. The observable deviations, as expected, are much lower than the theoretically derived relative underestimation of the tax liability if we refer to the mid-point of the class. It is remarkable that the high quality of the group simulation results arise when comparing not only the total tax revenues but also in almost every single class. The

sometimes substantial deviations found by other models in the lower and upper income classes (cf., e.g., Piketty and Saez, 2003:55, concerning the heterogeneity in the top income decile) are considerably reduced when we employ our discrete group simulation model. Moreover, the quality of the results of the discrete group simulation model is not dependent on the class limits chosen by the German Statistical Office. Even for simulations with tax scales whose basic tax-exempt amount does not correspond to the class limits set by the German Statistical Office, differences of similar structure and dimension occur, i.e. again very small deviations.

V.2 Tax base deductions simulation based on taxable income
It is desirable to find out whether the degree of precision of our group model obtained for tax scale simulations (Section V.1.) is achievable for the simulation of tax revenue effects caused by reforms of fixed (flat) amount tax base deductions as well. Unfortunately, no microsimulation was carried out by the German Statistical Office for this scenario, so comparison with our group simulation results is not possible. Instead, in the following we focus on the problem of tax deductions from the tax assessment base (c.f. O'Donoghue and Sutherland, 1999:576-577). In order to measure the fiscal impact of these deductions, their tax revenue effects are determined by considering a corresponding increase in the tax base within the simulation. Our conclusions can in principle be transferred to tax regulations that lead to an increase of the tax base and their tax revenue effects by simulating an adequate tax base reduction.

However, in case of such simulations the differences between micro and group analyses may increase if the underlying fixed amount is not deductible by all taxpayers and, further, if the (relative) distribution of the taxable income of the taxpayers who enjoy this deduction does not correspond to the (relative) distribution of the taxable income of all taxpayers. In order to improve the quality of the results of our group model, information about the distribution of the taxpayers enjoying this fixed tax privilege, as far as this information is available, should be considered explicitly in the simulation. From the published income tax statistics, as outlined already, the number of taxpayers and the total amount of fixed amount tax base deductions in thousands of deutschmarks per class is given. Therefore we have information about the distribution among different income classes, but not about the distribution of these amounts among the taxpayers within the classes. If the tables in the income tax statistics do not provide data on the taxable income of the taxpayers who benefit from this deduction, then for group simulation purposes we have to fall back on the sum of the taxable incomes of all taxpayers and hence, on the distribution of all taxpayers in this class derived from the group simulation. This may involve a larger, and possibly unacceptable, deviation from the results of a microsimulation.

Table 2 Tax scale based micro and group simulation of tax revenue for the basic tax scale (1995 income distribution)

TI class no.	taxable income (DM)		1990 tax scale			1996 tax scale
			income tax in (DM ' 000)		*relative difference (%)*	*relative difference (%)*
			microsimulation (German Statistical Office)	*group simulation*		
1	under 1		-	-	0.0000	0.0000
2	1 -	5,670	-	-	0.0000	0.0000
3	5,670 -	8,154	180,739	180,734	-0.0028	0.000
4	8,154 -	12,096	965,929	965,970	0.0042	0.0000
5	12,096 -	12,366	83,707	83,706	-0.0012	-0.0685
6	12,366 -	13,068	219,226	219,228	0.0009	0.0083
7	13,068 -	18,036	1,837,818	1,837,820	0.0001	0.0022
8	18,036 -	25,002	4,018,826	4,018,823	-0.0001	0.0008
9	25,002 -	30,023	4,408,774	4,408,747	-0.0006	-0.0003
10	30,023 -	40,013	14,380,662	14,381,434	0.0054	0.0034
11	40,013 -	50,004	15,660,133	15,660,329	0.0013	0.0010
12	50,004 -	55,728	7,173,713	7,173,689	-0.0003	-0.0003
13	55,728 -	58,644	3,108,599	3,108,590	-0.0003	-0.0003
14	58,644 -	60,048	1,403,312	1,403,312	0.0000	0.0001
15	60,048 -	66,366	5,465,481	5,465,480	0.0000	-0.0005
16	66,366 -	70,038	2,652,440	2,652,457	0.0006	0.0006
17	70,038 -	75,006	3,013,301	3,013,302	0.0000	0.0001
18	75,006 -	100,008	8,784,647	8,782,625	-0.0230	-0.0230
19	100,008 -	120,042	3,299,181	3,298,992	-0.0057	-0.0058
20	120,042 -	240,084	6,458,174	6,458,172	0.0000	0.0000
21	240,084 -	480,168	3,488,455	3,488,455	0.0000	0.0003
22	480,168 - 1,000,026		2,549,395	2,549,390	-0.0002	0.0000
23	1,000,026 or more		7,226,559	7,226,536	-0.0003	-0.0003
Total (ALL BANDS)			**96,379,068**	**96,377,792**	**-0.0013**	**-0.0018**

Source: German Statistical Office, Wiesbaden; own calculations.

Using the symbols defined in section IV.1 the problem can be presented formally as follows. From the aggregated data of the income tax statistics we know for each class *i* the frequency g_i of the existing tax facts (number of taxpayers, who are affected by this fact) and the sum of its value, G_i, where the average value of a class is given by

$$\overline{G_i} = \frac{G_i}{g_i}.$$

In the case of a fixed tax base deduction G_i is constant for each class. The financial consequences of this tax rule per class arise from the difference, ΔT_i, between the respective sum of the income tax of the class both including the effects of the deduction (T_i^g) and excluding its effect (T_i):

$$\Delta T_i = T_i^g - T_i,$$

where

$$T_i^g = h_{(TI)}^g \, t_{(TI+\overline{G_i})}.$$

Here $h_{(TI)}^g$ is the number of taxpayers with a specific *TI* who are affected by g_i.

Furthermore, $t_{(TI+\overline{G_i})}$ denotes the income tax for

the tax base *TI* which is increased by $\overline{G_i}$.

The degree of precision of the simulation is also influenced by whether or not we are informed about the sum of the taxable incomes of the taxpayers for the class *i* who deducted an amount (TI_i^g) due to special fixed tax regulations. Determining u_i and $h_{(TI)}^g$ using the equations (1.4) and (1.5) it is important whether we refer to the taxable income of all taxpayers (*TI)* or to the taxable income (TI_i^g) of those taxpayers who enjoy tax privileges and thus are included in g_i. If TI_i^g is known, then $h_i^g = g_i$. Otherwise u_i has to be determined on basis of TI_i and, for reasons of simplicity, we set

$$h_{(TI)}^g = h_{(TI)} \frac{g_i}{h_i}.$$

Proceeding like this, an identical distribution of the taxpayers with a specific taxable income for the respective class is assumed for all examined tax facts.

Precision is further reduced when applying a discrete group simulation model to determine tax revenue effects caused by tax base deductions that vary between taxpayers. This is imaginable in the cases of, for example, depreciation and loss offset allowances.

Since the actual distribution of taxpayers cannot be determined from the aggregated data we need appropriate assumptions on the distribution of the underlying tax deductions in each class analogous to those made with respect to the distribution of taxable income. These assumptions are necessary even if the distribution of taxpayers, the deductible amount in each income classes and even the sum of the taxable incomes of the taxpayers in question can be taken from the tables of the German income tax statistics. For our analysis, again for reasons of simplicity, we assume a uniform distribution so that for every taxpayer of a given class the average value

$$\overline{G_i} = \frac{G_i}{g_i}$$

that can be deduced from the sum of tax deductions of each class is taken as a proxy for the individual amount. Use of a class-specific average is is preferable to deducting the same fixed fixed amount (the overall average tax deduction) regardless of class.

The results of the microsimulation by the German Statistical Office on income tax revenue effects in case of limited loss offset are compared with those of our discrete group model in Table 3, using the same taxable income classes as Table 2. We analyze loss offset restriction as losses could not be compensated with positive earnings from other sources. In line with the comparison in Table 2 we apply the basic 1990 tax scale to determine the income tax. Applying the 1990 rather than the 1996 tax scale allows us to test whether the degree of modelling accuracy is independent of the class borders chosen by the German Statistical Office for a particular tax year.

The relative divergence of the income tax calculated on the basis of the group simulation and the income tax calculated on the basis of the microsimulation is presented in Table 3 for each income class as well as for all taxpayers. Furthermore, we distinguish between the basic and the splitting tax scale. Table 3 also includes relative differences in simulated financial consequences. The financial consequences are based upon the sum of the income tax of all taxpayers with negative earnings, in the case of either a complete or limited loss offset (cf. Wagstaff and van Doorslaer, 2001:307). The relative difference between the financial

Table 3 Tax base based micro and group simulation of tax revenue and the financial effects using *TI* tables in case of vertical loss offset restriction (1995 income distribution)

TI class no.	Percentage difference in the results of microsimulation (German Statistical Office) relative to the discrete group model			
	interpolation by TI_i^g		interpolation by TI_i	
	basic tax scale	splitting tax scale	basic tax scale	splitting tax scale
1	-1.4182	-100.0000	194.1274	263.1967
2	-34.5701	-35.6779	-37.4919	-38.2353
3	-20.3230	-17.8262	-19.9858	-17.7484
4	-14.7191	-11.8263	-15.0016	-11.6785
5	-9.4212	-7.9040	-9.4722	-7.8956
6	-11.7048	-7.3189	-11.7094	-7.3079
7	-9.8775	-6.5023	-10.5133	-6.4304
8	-6.3505	-4.3696	-6.3986	-4.9910
9	-5.1233	-2.9831	-5.2855	-3.0239
10	-3.6476	-2.7450	-3.6717	-3.1454
11	-3.3690	-2.6741	-3.8018	-3.1458
12	-2.4027	-2.4326	-2.7062	-2.4803
13	-2.7901	-2.3300	-2.8571	-2.3755
14	-2.3323	-2.4959	-2.4071	-2.5000
15	-2.6210	-2.4570	-2.7292	-2.5696
16	-2.5111	-2.2322	-2.7662	-2.2305
17	-2.0388	-2.1054	-2.1141	-2.1570
18	-1.6480	-1.7187	-2.6890	-2.5163
19	-0.1687	-0.1625	-0.5135	-0.3792
20	-0.1687	-0.0181	-2.3527	-0.7592
21	-0.0071	-0.0080	-0.5272	-0.1134
22	-0.0038	-0.0037	0.0445	0.2148
23	-0.0010	-0.0010	-5.2218	0.0058
total	**-1.9124**	**-3.7316**	**2.6085**	**3.0275**
total without class 1	**1.9282**	**-1.8186**	**-3.5040**	**-2.1424**
financial effects	**-6.8567**	**-14.5591**	**14.6799**	**13.0822**
financial effects without class 1	**-7.5351**	**-7.4848**	**-7.7054**	**-7.6267**

Source: German Statistical Office, Wiesbaden; own calculations.

consequences of a refusal to allow vertical loss offset is shown at the end of the table.

In addition, the group simulation was carried out on the basis of two differently aggregated data sets. The first group simulation is based on tabulated data from the sample projected by the German Statistical Office. This sample contains data for taxpayers with a negative income, i.e. the sum of the taxable income of these taxpayers per class is known (TI_i^g). This group specific information cannot be found in the publicly available model results. Rather, it was prepared by the German Statistical Office as a special statistical evaluation for this research project only. In contrast, the second group simulation used the sum of the taxable income of all taxpayers of the class (TI_i) provided in the tabulated data to simulate the distribution of the tax bases within the class. The relevant details for all taxpayers are included in the published statistics.

Concentrating on the tax revenue effects of tax base deductions (which may be different for every taxpayer), a comparison of the results of Tables 2 and 3 shows that the deviations of group simulation results from those of the microsimulation model are substantially greater than those found when simulating different tax scales. When interpolating using the class sum of the taxable income of the taxpayers with a negative income, TI_i^g, we find that the group simulation results are lower, for all taxable income classes, than the microsimulation model results (negative relative differences). The total effect, across all taxable income classes, is a deviation -1.9% (basic tax scale) and -3.7% (splitting tax scale). When interpolating using TI_i, the class sum of the taxable income of all taxpayers of the class, the group simulation results are once again consistently lower than for the microsimulation model, with the notable exception of very highest and lowest taxable income bands. These apparently minor differences, however, have a significant impact. The overall net deviation, summed across all classes, becomes positive (rather than negative), with positive deviations of 2.6% (basic tax scale) and 3.0% (splitting tax scale).

The differences are largest in the lower income classes and decrease as the tax base increases. The greatest relative difference is observed for class 1, which includes taxpayers with a taxable income less than one DM. Since this class is not further subdivided in the income tax statistics but covers a wide range of negative taxable incomes, here the group simulation model is highly inaccurate. As a consequence, estimating the number of taxpayers with positive income greater than the basic tax-exempt amount due to vertical loss offset restriction is rather unreliable. Besides, the results in this class depend on the lower class boundary which must be determined heuristically. Including the class of the taxpayers with a taxable

income less than one DM is reasonable only for microsimulation of tax revenue effects if we want to analyze an increase in the tax base - as far as these taxpayers are affected by it.[10] Due to the lack of data, in this case a group model can only arbitrarily lead to similar results as a microsimulation. If the class of taxpayers with a taxable income less than one DM is neglected in simulation, comparing micro and group models leads to relative deviations in tax revenues for all taxpayers with a negative income employing interpolation using TI_i^g of 1.9% (basic tax scale) and -1.8% (splitting tax scale) and further, using TI_i of -3.5% (basic tax scale) and -2.1% (splitting tax scale).

We realize that the tax revenue calculated by microsimulation for the unmodified tax base (Table 2) does not differ as much as from the one determined by group simulation as do the tax revenues assuming a modified tax base (Table 3). (The modified taxable income is given by the taxable income increased – for example by losses that have not yet been offset against profits.) Therefore, the financial consequences of the tax base modification invoke substantially greater relative deviations between the microsimulation and the group simulation. The differences occurring in the lower income classes particularly preponderate. The relative deviations between the microsimulation and the group simulation for the overall financial effects including all income classes are -6.9% (basic tax scale) and -14.6% (splitting tax scale) using TI_i^g and are 14.7% (basic tax scale) and 13.1% (splitting tax scale) referring to TI_i. If we neglect the lowest income class, relative deviations of about -7.5 % (basic and splitting tax scale) arise in the context of the interpolation of TI_i^g -7.7 % (basic tax scale) and -7.6% (splitting tax scale) can be found by employing TI_i. Obviously, a group simulation excluding the inaccurate values of the first class leads in principle to an underestimation of the financial effects. This finding meets the expectations since by relying on the average amount of tax base deductions per corresponding taxpayer we determine the lower boundary of the possible tax revenue shortfall.

Furthermore, Table 3 clarifies that the results of the group simulation that are based on the class sum of the taxable income of the taxpayers with a negative income (TI_i^g) involve – as expected – a tendency towards fewer deviations from the microsimulation results than is the case in a simulation that refers to the class sum of the taxable income of all taxpayers of the class (TI_i). From this, we cannot conclude that the structure of deviation identified here will generally be observable because the (unknown) distribution of the taxpayers within a class in principle may differ by class and by the examined tax facts. This is clarified comparing the class specific results in Table 3.

V.3 Tax scale simulation based on total amount of income

Most of the tables provided by the German Statistical Office on income tax, in particular those relating to specific tax rules, are not arranged according to size of the taxable income but rather to size classes of the "total amount of income". Of course, taxable income would be a better group attribute for the underlying research question. Therefore, and in general, it would be desirable that the Statistical Office releases tables of this type. This would improve group model results and thus tax effect simulations. On the other hand, the Statistical Office already releases tables covering more than 1000 different attributes, and it is clearly not possible to produce a set of tables sufficient to satisfy all possible research questions in advance. In addition there are known problems in defining total taxable income (cf. O'Donoghue and Sutherland, 1999; Goolsbee 2000). In the future the fiscal authorities might perhaps make it easier to obtain specially commissioned tables, in so far as this is possible without compromising on respondent confidentiality. In the meantime, given the lack of data on taxable income, we have to make use of the data provided on "total amount of income".

Once again analyzing taxpayers that are subject to either the basic or splitting tax scales, the database can be described formally as follows. The supplied data provide a categorization per total amount of income for $j=1$ to m classes with class borders $[c_j, d_j]$, where $c_1=-\infty$, $d_1=-1$, $c_2=0$ and $d_m=\infty$. For classes $j>1$ the taxpayers have a taxable income greater than zero DM. The first class ($j=1$) contains the so called cases of loss which occur if the taxpayer has an assessed negative income. A negative value can result when determining of the sum of the earnings from different sources of income or, later in the assessment pattern, when determining the taxable income, for example due to the deduction of extra expenditures and extraordinary expenses.

For each class j we know:
- the frequency h_j in class j (number of taxpayers in the class for whom a taxable income has been assessed),
- the frequency g_j of a tax fact (number of taxpayers who meet this fact) and the value G_j of this tax fact (in thousands of DM or €),
- the sum of the taxable incomes of all taxpayers in this class TI_j and
- the sum of the assessed income tax of all taxpayers in this class T_j.

Applying the group simulation model to data from tables that are arranged according to total amount of income (*TAI*) the following problem arises. The distribution of taxpayers with a specific taxable income ($h_{(TI)}$) is difficult to estimate due to the fact that for the taxpayers of a *TAI*-class j only the average taxable income of the class,

$$\overline{TI}_j = \frac{TI_j}{h_j} \; ,$$

can be determined directly. The interval range $[a_i, b_i]$ of the possible taxable income of these taxpayers cannot be deduced from the *TAI* tables.

By mapping a taxpayer to a certain *TAI* class we can only determine the upper limit of the taxable income b_i as the theoretical maximum taxable income of the class by reducing the upper limit of the *TAI* class d_j by the minimum fiscal reductions, for example allowances for special expenses. In contrast, a theoretical lower limit for the taxable income a_i cannot be determined because the taxable income can adopt any value below the upper bound of the *TAI* class d_j due to various discounts on the total amount of income, for example special expenses, loss offset or extraordinary expenditures. Consequently, in this case the lower interval limit a_i (smallest possible taxable income) must be estimated roughly, implying relatively high inaccuracy of the results of simulation. In order to reduce the deviations in group simulation caused by this deficit of information cross tables were provided by the German Statistical Office for our analysis. These cross tables allow us to restructure part of the aggregated data of the income tax statistics that are grouped according to total amount of income (*TAI)* and rearrange them according to classes of taxable income (*TI)*.

In these cross tables the absolute frequency of the taxpayers, h_i, with a taxable income in class i and the sum of the taxable income $TI_{i,}$, are brought together with the absolute frequency of the taxpayers, h_j, with a total amount of income in class j and the sum of the taxable incomes of these taxpayers, TI_j. As a result, we obtain a matrix of the absolute frequencies of the taxpayers, h_{ij}, and the necessary sums of the taxable income, TI_{ij}.

Using this matrix it is possible to estimate the distribution of the taxpayers with a specific taxable income from the aggregated data of the income tax statistics grouped according to the class attribute "total amount of income", as in eqns. (1.4) and (1.5):

$$(1.8) \quad h_{(TI)} = \frac{h_{ij}}{(b_i - a_i + 1)} + \frac{TI - a_i - \left(\dfrac{b_i - a_i}{2}\right)}{\left(\dfrac{b_i - a_i}{2}\right)} u_{ij},$$

where

$$(1.9) \qquad u_{ij} = \frac{TI_{ij} - \dfrac{a_i + b_i}{2} h_{ij}}{\displaystyle\sum_{TI=a_i}^{b_i} \dfrac{TI - a_i}{\dfrac{b_i - a_i}{2}} \dfrac{\dfrac{b_i - a_i}{2}}{\dfrac{b_i - a_i}{2}}} TI \, .$$

Then, employing the discrete group simulation model the income tax revenues can be determined by equation (1.3).

A comparison of the results of the group simulation based on the *TAI* tables with those of the microsimulation shows that applying cross tables (prepared by the German Statistical Office) to tax scale simulation with the discrete group model provides fairly accurate results. The relative deviations between the group simulation and microsimulation results are shown in Table 4. The simulation was conducted using the 1990 tax scale. For those taxpayers who are taxed at the basic rate group simulation leads to a minor overestimation of the income tax, with the estimate of total income tax paid being 0.027% more than that provided via microsimulation. In the case of the splitting tax scale the results of the group simulation differ by -0.048%; i.e. they are slightly lower than for microsimulation. The

minor overestimation for the basic tax rate is caused by the underlying data of the tax base.

In Table 4 a comparison of the data based on the respective simulation of the tax base precedes the comparison of the aggregated income tax. This comparison clarifies to what extent the data extrapolated from the sample, as used for the microsimulation, differ from the data from the income tax statistics used in the group simulation. As Table 4 shows, the differences between the results produced using the sample provided by the German Statistical Office and the values for the basic population contained in the published income tax statistics are very small. This suggests, therefore, that the simulation results are barely affected by the structural differences in the datasets.

In the group simulation based on the *TAI* tables the fixed and variable reductions and discounts from the tax base are considered in line with the procedure in sections V.1 and V.2. The tax revenue that would result without taking account of specific tax facts (*g, G*) can be estimated analogously to eqn. (1.7). Here, we usually assume

$$h_{(TI)}^{g} = h_{(TI)} \frac{g_j}{h_j}$$

Table 4 Tax scale based micro and group simulation of tax revenue using a sample of the original microdata and grouped data of *TAI* tables (1995 income distribution)

TI class no.	Percentage difference in results reported					
	(i) by sample relative to *TAI* tables				(ii) by group simulation relative to microsimulation	
	number of taxpayers		taxable income		calculated income tax	
	basic tax scale	splitting tax scale	basic tax scale	splitting tax scale	basic tax scale	splitting tax scale
1	1.4916	-0.0093	0.0934	0.0690	0.0000	0.0000
2	0.8065	-0.000	0.4750	0.0010	0.0000	0.0000
3	0.0927	-0.0020	0.0891	-0.0023	0.0702	-1.0534
4	0.0692	-0.0005	0.0687	-0.0009	0.0723	-0.2947
5	-0.0458	0.0226	-0.0451	0.0233	-0.0455	-0.1903
6	0.0984	-0.0266	0.0993	-0.0269	0.1007	-0.2267
7	0.0482	0.0040	0.0493	0.0036	0.0502	-0.1414
8	0.0784	-0.0015	0.0780	-0.0016	0.0777	-0.0970
9	0.0649	0.0030	0.0659	0.0030	0.0657	-0.0680
10	0.0386	-0.0010	0.0391	-0.0008	0.0447	-0.0544
11	0.0324	0.0001	0.0325	-0.0002	0.0338	-0.0445
12	0.0332	-0.0019	0.0333	-0.0023	0.0329	-0.0392
13	0.0073	0.0176	0.0073	0.0171	0.0070	-0.0172
14	0.0788	0.0000	0.0790	0.0005	0.0792	-0.0323
15	0.0212	0.0000	0.0216	0.0001	0.0217	-0.0311
16	-0.0373	0.0077	-0.0384	0.0084	-0.0383	-0.0203
17	0.0761	-0.0106	0.0754	-0.0105	0.0751	-0.0378
18	0.0235	0.0019	0.0228	0.0021	-0.0006	-0.0384
19	0.0064	-0.0073	0.0057	-0.0075	-0.0004	-0.0302
20	-0.0046	-0.0094	-0.0130	-0.0072	-0.0162	-0.0180
21	0.0467	0.0000	0.0369	0.0000	0.0357	-0.0048
22	0.0000	0.0000	0.0000	0.0000	0.0000	-0.0021
23	0.0000	0.0000	0.0000	0.0000	0.0001	-0.0003
total	**0.2782**	**-0.0007**	**0.0388**	**-0.0020**	**0.0269**	**-0.0477**

Source: German Statistical Office, Wiesbaden; own calculations

because the taxable income (TI_j^g) of the taxpayers who are subject to such tax base deductions is not known. This implies that an identical distribution of taxpayers with a specific taxable income is assumed for each *TAI* class. This simplifying procedure may lead to greater deviations when we are using data that is classified according to the total amount of income (*TAI*), as is the case for the group simulation, in comparison to data from tables that are arranged according to taxable income (*TI*). This is true since the assumption of an identical distribution of taxpayers with a specific taxable income for an interval of the taxable income [a_i, b_i] leads to smaller differences than for an interval of the total amount of income [c_j, d_j].

V.4 Tax base deductions simulation based on total amount of income

The results in Table 4 are in line with the corresponding findings presented in Table 5, again based on data from *TAI* tables, regarding the effects of vertical loss offset restrictions on tax revenues. In the case of a group simulation based on the *TAI* tables the tax bases before the loss simulation already differ from those of the microsimulation because of the different under-

lying datasets.

This is caused by the fact that the taxable income of the taxpayers with a negative income, TI_j^g, cannot be derived from the *TAI* tables. The tax base of the taxpayers who obtained a negative income must be estimated instead using the taxable income of all taxpayers of the respective class, TI_j. The number of taxpayers with a negative income, g_j, and the sum of the negative income per class, G_j, are given in the *TAI* tables. Therefore, we realize only slight differences in the total sum of the negative incomes: -0.17% for the basic rate taxpayers and -0.35% for the splitting scale taxpayers.

For individual classes it turns out that the transition of negative income *TAI* classes into *TI* tables can lead to severe deviations in the individual classes. This is due to the fact that running a group simulation based on *TAI* tables the distribution of the average amount of the negative income in a *TAI* class, G_j, is made according to the distribution of the taxable income of all taxpayers of this class. Consequently, the taxable income and the modified taxable income in each single class may also show strong

Table 5 Tax base based micro and group simulation of tax revenue and the financial effects using a sample of the original microdata and grouped data of *TAI* tables in the case of vertical loss offset restrictions (1995 income distribution)

TI class No.	Percentage deviation of the microsimulation results of the German Statistical Office from those of the discrete group model							
	negative income		taxable income		modified taxable income		calculated income tax on modified taxable income	
	basic tax scale	splitting tax scale	basic tax scale	splitting tax scale	basic tax scale	splitting tax scale	basic tax scale	splitting tax scale
1	-0.95	-1.34	-75.05	-63.24	73.73	147.02	189.31	136.41
2	-50.78	-37.57	-31.18	-18.79	-46.11	-31.99	-68.99	-59.58
3	-13.07	-11.43	0.06	-1.08	-7.96	-6.68	-29.48	-26.05
4	9.25	-1.37	1.27	1.43	11.13	0.25	-5.83	-12.02
5	17.25	14.01	19.80	8.50	18.53	10.45	6.81	2.77
6	29.06	13.51	18.99	8.29	23.66	10.09	11.29	3.09
7	8.07	9.75	6.03	8.85	6.91	9.20	-3.36	2.21
8	-7.58	0.06	-4.93	2.19	-5.88	1.68	-12.15	-3.06
9	-11.46	5.50	-3.57	0.41	-5.95	1.39	-11.73	-1.15
10	9.25	16.26	8.10	6.99	8.37	8.56	4.61	6.37
11	1.33	58.41	10.69	23.48	8.74	29.04	4.37	28.36
12	1.42	40.82	0.07	14.03	0.31	18.20	-1.98	17.20
13	22.78	20.50	16.16	4.24	17.33	6.74	14.63	5.40
14	5.67	12.35	11.61	0.79	10.48	2.56	7.42	0.83
15	8.27	-5.75	4.16	-6.05	4.87	-6.00	2.51	-8.32
16	-0.56	-23.20	1.82	-16.78	1.38	-17.81	-1.34	-20.12
17	28.67	-30.56	13.76	-22.04	16.32	-23.39	15.32	-25.64
18	21.9	-55.51	5.5	-33.95	8.49	-37.79	9.1	-40.88
19	36.01	-32.40	16.81	-26.38	20.69	-27.50	22.11	-27.81
20	-24.27	1.15	-17.82	-0.42	-19.11	-0.15	-19.82	-0.25
21	75.79	9.33	3.77	0.89	16.45	2.18	17.94	2.33
22	-2.36	47.36	-1.76	0.51	-1.85	7.03	-1.85	7.42
23	-68.43	-40.02	-15.47	1.35	-20.20	-2.53	-20.31	-2.59
total	**-0.17**	**-0.35**	**19.30**	**5.31**	**26.87**	**9.74**	**2.80**	**-1.21**
total without class 1	**0.55**	**0.18**	**0.25**	**-1.20**	**21.85**	**5.37**	**-3.15**	**-3.95**
financial effects							**16.47**	**2.31**
financial effects without class 1							**-5.09**	**-8.79**

Source: German Statistical Office, Wiesbaden; own calculations.

deviations. Here, however, the unmodified taxable income neglecting the class of the taxpayers with a taxable income less than one DM can be determined relatively exactly. We find a deviation of 0.25% applying the basic tax scale and of -1.2% applying the splitting tax scale.

On this basis we can reconcile the findings regarding the tax revenue effects of reduced loss offset allowances produced via group simulation on the basis of *TI* tables (see Table 3) and via group simulation on the basis of *TAI* tables (Table 5). For the class of taxpayers with a taxable income of less than one DM only very inaccurate results can be obtained. Consequently, this leads again to an overestimation when determining the total financial effects of reduced loss offset using the group model. If we exclude the class of taxpayers with negative income from the analysis we receive an underestimation of the financial effects of 5.1% in the case of the basic tax scale and 8.8% in the case of the splitting tax scale. These deviations are similar to the differences realized by applying *TI* tables (c.f. Table 3).

VI SUMMARY

In this paper we compare the results obtained by microsimulation with those generated by a discrete group model using differently classified data. Through this comparison we point out and quantify the possible effects of the simplified procedure of the group model, as well as the loss of information involved in using aggregated and incomplete data. The differences identified by concentrating on specific examples do not provide generally validated values. Nevertheless, they indicate the magnitude of possible inaccuracies caused by a group simulation. We find that group simulation under certain circumstances provides results very close in accuracy to those obtained via microsimulation. Furthermore, for those cases in which group simulation is the appropriate tool, we provide a very simple method to interpolate the income distribution and thereby the tax distribution within the classes. This interpolation makes future estimates of tax revenues a lot easier. These results are interesting and important as microsimulation is far more time consuming and resource intensive than group simulation, whilst for cross-country and time series analyses microdata are not usually available and we may well have to fall back, in any case, on group models.

Summarizing, we find that applying the group simulation model to analyze tax scale effects leads to very good results. The differences between the results derived by microsimulation in comparison to group simulation increase if we determine the financial effects of modified tax bases, particularly if tax base cuts vary between taxpayers and if we take account of the class of

the taxpayers with a taxable income of less than one DM. Neglecting this class we identify a systematic underestimation simulating the financial consequences of a modified tax base with the group model, assuming a progressive tax scale. In this situation, as disaggregated data on the tax base modification is not available we have to adjust the empirical class average of the tax base reduction. If the group simulation data is not arranged according to the taxable income but rather to the total amount of income we tend to find greater deviations from the microsimulation results in sum as well as per class.

From this we can conclude that, if the input data are sufficiently detailed and complete, micromodels will always be superior in accuracy and provide a more sophisticated tool for estimating tax revenue effects. If, on the other hand, microsimulation relies upon a sample – as is the case for the German Statistical Office model – and not on a microdataset with complete coverage, very large deviations may result. These deviations arise from the structure of the stratified random sample, which will not always be representative of circumstances and facts due to a relatively small frequency of responses in each category. In this case, the possible errors due to the group model and the aggregated database are considerably smaller than in case of a microsimulation because in the group simulation we apply data that is based on the overall population in the income tax statistics.

Aiming to determine and analyze the tax revenue effects of alternative tax settings, in particular the financial effects of specific fiscal regulations, the group simulation model introduced here can offer a good compromise between a) allowing the model and the data to reflect a complex situation as accurately as possible and b) the possible accuracy of a model that is based on limited resources and data. To further improve the utility of group models for tax effect simulations when access to microdata is restricted, we urge the fiscal authorities to publish, or provide a mechanism for commissioning, additional statistical tables on a taxable income rather than total amount of income basis.

Notes

[1]   If microsimulation is indispensable for certain complex research questions and there is a lack of tax microdata it might be recommendable to run a microsimulation on the basis of synthetic microdata generated from grouped data. These synthetic data could be generated using various data bases, for example data from other samples (consumer panel data) or from national accounting and merge these data with the grouped data from the tax statistics. Users

of such synthetic microdata have to check whether the generated microdata are representative for the underlying research question and taxpayers. In case of analyses on tax revenue effects synthetic microdata will usually not be representative. If microdata is not representative, for example, because of missing important attributes in the aggregate data base, the results of a microsimulation will be wrong. For other research questions this data inaccuracy can be less relevant and hence negligible.

[2] For a continuous-time approach cf., e.g., Galler (1997).

[3] In accordance with § 32 para. 2 EStG, which describes the German income tax scale, the income tax scale only has to be applied to full DM (deutschmark) or euro amounts.

[4] A differing income tax may result from applying the "exemption with progression" rule or specific tax rates for extraordinary earnings.

[5] This is also valid for the attribute "tax scale income tax" published in the income tax statistics since the tax scale induced income tax is influenced by special rate prescriptions as well.

[6] Here, in particular, the transition from the zero-zone of the tax schedule that is determined by the basic tax-exempt amount to the next zone is problematic, since in the case of an average taxable income of the class lying below the basic tax-exempt amount the aggregated income tax of the class would be zero.

[7] The sign of the difference of the number of taxpayers at the class border and the one in the mid-point of the class is determined by the position of the average taxable income of the class

$$\overline{TI_i}$$

in relation to the mid-point of the class

$$\frac{a_i + b_i}{2}.$$

If $\overline{TI_i} > \frac{a_i + b_i}{2}$, then

$$h_{(b_i)} - h_{\frac{a_i+b_i}{2}} = 1 \text{ and } h_{(a_i)} - h_{\frac{a_i+b_i}{2}} = 1.$$

Whereas if $\overline{TI_i} < \frac{(a_i + b_i)}{2}$ the differences are given by $h_{(a_i)} - h_{\left(\frac{a_i+b_i}{2}\right)} = 1$ and $h_{(b_i)} - h_{\left(\frac{a_i+b_i}{2}\right)} = -1$.

[8] Due to this simplifying procedure the modelled number of taxpayers with a specific income, $h_{(TI)}$, is not necessarily integer.

[9] In contrast, Piketty and Saez (2003) employ group data for high income taxpayers and assume a Pareto distribution. They do no simulate taxation and have a different objective. Piketty and Saez (2003) analyze the income distribution and income composition and further the tax burden for high income

taxpayers. Instead, we focus on tax revenue effects from any income. Nevertheless, the tax revenues estimated for high income taxpayers by either their or our model would hardly deviate due to the underlying income distributions. To analyze tax revenue effects, our approach incorporates the advantage of simplicity of the interpolated income distribution. Moreover, in contrast to the Pareto distribution, this distribution can be applied to low, medium as well as high taxable income without sacrificing accuracy in estimation results.

[10] Several studies solely consider taxpayers or households with positive income. Cf., e.g., Zandvakili (1994:479).

REFERENCES

Atkinson A B (2007) 'The distribution of top incomes in the United Kingdom 1908-2000', in Atkinson A B and T Piketty (Eds), *Top Incomes over the Twenties Century. A Contrast between Continental European and English-Speaking Countries*, Oxford: Oxford University Press, 82-140.

Atkinson A B and Bourguignon F (Eds.) (2000) *Handbook of Income Distribution*, vol. 1, Amsterdam: North-Holland.

Berglas E (1971) 'Income Tax and the Distribution of Income. An International Comparison', *Public Finance,* 24(4), 532-545.

Bork C and Petersen H (2000) 'Revenue and distributional effects of the current tax reform proposals in Germany - An Evaluation by Microsimulation', in Petersen H-G and Gallagher P (Eds.), *Tax and Transfer Reform in Australia and Germany,* Berlin: Berliner Debatte Wissenschaftsverlag, 219-235.

Caldwell S B (1985) 'Synthesis via microsimulation', in Nakamura A and Nakamura M. (Eds.), *Synthesis through Microanalytic Simulation*, New York: Elsevier Science Publishers, 71-75.

Callan T and Sutherland H (1997) 'The impact of comparable policies in European Countries: Microsimulation Approaches', *European Economic Review,* 41(3-5), 627-633.

Citro F C and Hanushek E A (1991) *Improving Information for Social Policy Decisions,* Washington, D.C.: National Academy Press.

Cowell F A (1984) 'The structure of American income inequality', *Review of Income and Wealth,* 30(3), 351-375.

Dell F (2007) 'Top incomes in Germany throughout the Twentieth Century: 1891-1998', in Atkinson A B and T Piketty (Eds.), *Top Incomes over the Twentieth Century: A Contrast Between European and English Speaking Countries,* Oxford: Oxford University Press, 365-425.

Galler H P (1997) 'Discrete-time and continuous-time Approaches to dynamic microsimulation reconsidered', Technical Paper 13, University of Canberra, Canberra.

Goolsbee A (2000) 'What happens when you tax the rich? evidence from excecutive compensation', *Journal of Political Economy,* 108 (2), 352-378.

Haan P and Steiner V (2005) 'Distributional effects of the German Tax Reform 2000 - A behavioral microsimulation analysis', *Schmollers Jahrbuch - Journal of Applied Social Science Studies,* 125(1), 39-49.

Kakwani N C (1977) 'Measurement of tax progressivity: an international comparison', *The Economic Journal,* 87(345), 71-80.

Kraus F (1981) 'The historical development of income inequality in Western Europe and the United States', *The development of welfare states in Europe and America,* New Brunswick, NJ: Transaction Books and the HIWED Project.

Loizides I (1988) 'The decomposition of progressivity indices with applications to the Greek taxation system', *Public Finance,* 43 (2), 236-247.

McClung N (1986): 'The art of transfer policy analysis', in Orcutt G, Merz J and Quinke H. (Eds.), *Microanalytic Simulation Models to Support Social and Financial Policy*, Amsterdam et al.: North-Holland, 101-112.

Merz J (2000) 'The distribution of income of self-employed, entrepreneurs and professions as revealed from micro income tax statistics in Germany', in Hauser R and Becker I (Eds.), *The Personal Distribution of Income in an International Perspective,* Berlin: Springer, 99-128.

Morrisson C (2000) 'Historical perspectives on income distribution: the case of Europe', *Handbook of income distribution,* Paris: Elsevier Science B.V., 217-260.

O´Donoghue C and Sutherland H (1999) 'Accounting for the family in European income tax systems', *Cambridge Journal of Economics,* 23(5), 565-598.

Orcutt G H (1957) 'A new type of socio-economic system', *The Review of Economics and Statistics,* 39(2), 116-123.

Orcutt G H (1982) 'Microanalytic system simulation', in Nakamura A. and Nakamura M. (Eds.), Synthesis through Microanalytic Simulation, New York: Elsevier Science Publishers, 1-8.

Orcutt G, Merz J and Quinke H (1986) *Microanalytic Simulation Models to Support Social and Financial Policy,* Amsterdam: North Holland.

Piketty T (2003) 'Income inequality in France, 1901-1998', *Journal of Political Economy,* 111 (5), 1004-1042.

Piketty T and Saez E (2003) 'Income inequality in the United States, 1913-1998', *The Quarterly Journal of Economics,* 118 (1), 1-39.

Piketty T and Saez E (2007) 'How progressive is the US federal tax system? A historical and international perspective', *Journal of Economic Perspectives*, 21(1), 3-24.

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 (3), 327-365.

Saez E and Veall M (2007) 'The evolution of high incomes in Canada, 1920-2000', in Atkinson A B and T Piketty (Eds), *Top Incomes over the Twenties Century. A Contrast between Continental European and English-Speaking Countries*, Oxford: Oxford University Press, 226-308.

Sutherland H (1995) 'Static microsimulation models in Europe - A Survey', DAE Working Paper No. 9523, University of Cambridge.

Wagenhals G (2001) 'Incentive and redistribution effects of the German tax reform 2000', *Finanzarchiv,* 57(3), 316-332.

Wagenhals G (2004) 'Tax-benefit microsimulation models for Germany: A Survey', Hohenheimer Diskussionsbeiträge Nr. 235, Institut für Volkswirtschaftslehre, University of Hohenheim.

Wagstaff A, van Doorslaer E V, van der Burg H et al. (1999) 'Redistributive effect, progressivity and differential tax treatment: personal income taxes in twelve OECD countries', *Journal of Public Economics,* 72 (1), 73-98.

Wagstaff A and van Doorslaer E V (2001) 'what makes the personal income tax progressive? A comparative analysis for fifteen OECD countries', *International Tax and Public Finance,* 8(3), 299-315.

Zandvakili S (1994) 'Income distribution and redistribution through taxation: an international comparison', *Empirical Economics,* 19(3), 473-491.

Zwick M (2001) 'Individual tax statistic data and their evaluation possibilities for the scientific community', *Schmollers Jahrbuch - Journal of Applied Social Science Studies,* 121 (4), 639-648.