DEVELOPMENT OF A NEW INCOME STRATIFIER FOR A SAMPLE OF INDIVIDUAL TAX RETURNS

John L. Czajka, Mathematica Policy Research, Inc.
600 Maryland Ave., S.W., Washington, D.C. 20024

Each year the Statistics of Income (SOI) Division of the Internal Revenue Service (IRS) draws a sample of individual tax returns filed during that year. The principal usage of the individual sample lies in the production of aggregate statistics. Annually the SOI Division publishes estimated totals for about 200 income and tax items, distributed by adjusted gross income (AGI) class or by filing status. However, policy analysts in the Treasury Department, Congressional agencies, and elsewhere also use SOI microdata--primarily for research on the operation of the tax system. Applications include simulations at the individual level to estimate the revenue implications and distributional impact of prospective changes to the tax code. The broad scope and volume of tax legislation in recent years has placed extreme demands upon the individual tax return microdata, and key users have pressed for revisions to the design of the individual sample to improve its ability to support policy research needs. This paper focuses upon one aspect of the individual sample redesign--namely, the stratification by income.

Obviously, the SOI sample design must be able to support the estimates of income and tax aggregates up to acceptable levels of precision. In addition, to serve the policy modeling needs the sample design should include several other features. The design must provide adequate samples of returns from key subpopulations, including the range of income brackets and tax brackets, types of filers, and age groups. The design must also yield adequate samples of policy-relevant line items--e.g., capital gains, social security benefits, and the major itemized deductions. To support estimates of the incidence of particular tax changes, the representation of key line items must be strong across major subpopulations.

These multiple requirements must be addressed by the revised stratification, which is scheduled to be implemented with the 1989 tax year sample. This paper describes the current sample design, discusses its strengths and limitations, and reviews the prospective revisions that are under consideration. The paper then outlines the research effort that is being undertaken to develop a new income stratifier.

DESCRIPTION OF THE CURRENT STRATIFICATION

Each tax return processed by the IRS during a given calendar year is assigned to a sampling stratum and then subjected to SOI selection at a rate defined for that stratum. For most returns the stratification is based solely on income; each return is assigned to one of nine income classes. Returns with business income or loss (Schedule C) or farm income or loss (Schedule F) are stratified on a combination of income and total receipts. Other conditions may result in a return's being assigned to one of the specialized strata, which vary in number from year to year and which take precedence over the 27 basic strata. Table 1 lists the 33 strata employed in selecting the 1985 filing year sample (from returns processed in 1986). Population and sample counts plus the realized sampling rates are reported in the table as well.

Since 1982 the income stratification has been based on the larger absolute value of a positive amounts total (PAT) and a negative amounts total (NAT), calculated from the income components of the taxpayer's AGI. The components represent an exhaustive decomposition of the "total income" line appearing on the first page of the individual income tax return. The components have changed slightly over time with changes in the tax form. The current components are listed below:

- Wages, salaries, tips, etc.
- Taxable interest income
- Dividend income
- Taxable refunds of state and local income taxes
- Alimony received
- Business income or loss (+/-)
- Capital gain or loss (+/-)
- Other gains or losses (+/-)
- Taxable pensions, IRA distributions, annuities and rollovers
- Rental or royalty income or loss (+/-)
- Partnership and S corporation income or loss (+/-)
- Estate and trust income or loss (+/-)
- Income or loss from Real Estate Mortgage Investment Conduits (+/-)
- Windfall profit tax credit or refund
- Farm income or loss (+/-)
- Unemployment compensation
- Taxable amount of social security benefits
- Other income

Items with possible negative amounts are identified by (+/-) following the name.

Only positive amounts are included in PAT, and only negative amounts are counted in NAT. Therefore the larger of PAT or NAT will exceed the taxpayer's line item total if any of the components is negative. This separate summation of positive and negative components may result in a return being assigned to a higher income stratum than if the classification were based on net income.

The sampling rates utilized in selecting the 1985 SOI sample ranged from .02 percent in the lowest stratum for business returns, to 100 percent in two of the specialized strata and in the highest income strata for all types of returns.

CRITIQUE OF THE CURRENT INCOME STRATIFICATION

The SOI individual sample must serve two distinct and, to some degree, conflicting sets of needs. First, it must be able to support precise estimates of a large number of income and tax aggregates. Second, the sample must provide a microdata base for tax policy research. The conflict between these two sets of objectives is clear. The optimum design for a sample that will be used to estimate numerous income aggregates will include stratification by an appropriate measure of income, and sampling rates that increase sharply with the income level of the stratum. However, a sample that will be used to support policy research requires a different stratification and sample allocation.

The income stratification employed currently in the individual sample was designed to address the need for precise aggregate statistics. Except for the inclusion of specialized strata, plus business and farm strata, the design does not address the policy modeling needs. Moreover, even with regard to aggregate statistics the current sample design invites suggestions for improvement.

There are several dimensions to the specification of an income stratifier for a sample of individual tax returns. These include:

- o inclusiveness of the income concept
- o treatment of losses

Table 1--Number of Forms 1040, 1040A and 1040EZ in the Population and Sample, 1985

Description of the sample strata		Number of returns: Population count	Number of returns: Sample count	Realized sampling rate
Grand total		101,836,347	121,480	0.12%
Form 1040 returns only with adjusted gross income of $200,000 and over with no income tax after credits and no additional tax for tax preferences, total		943	943	100.00
Form 1040 returns only with combined Schedule C (business or professional) net profit or net loss of $200,000 and over, total		13,304	13,304	100.00
Larger of total income amounts or total loss amounts	and Size of business receipts plus farm receipts			
Forms 1040 only with Form 2555		166,883	125	0.07
Under $1,000,000	Under $20,000,000	166,801	78	0.05
$1,000,000 and over	Any amount	82	47	57.32
Under $1,000,000	$20,000,000 and over			
Forms 1040 only with Form 1116, but without Form 2555		428,910	1,993	0.46
Under $1,000,000	Under $20,000,000	425,108	100	0.07
$1,000,000 and over	Any amount	3,802	1,893	49.79
Under $1,000,000	$20,000,000 and over			
Forms 1040 only with Schedule C, but without a Form 2555 or Form 1116		12,255,095	24,850	0.20
Under $20,000	Under $200,000	4,661,111	3,106	0.07
$20,000 under $50,000	Under $500,000	5,182,441	4,345	0.08
Under $20,000	$200,000 under $500,000			
$50,000 under $100,000	Under $1,000,000	1,839,912	4,114	0.22
Under $50,000	$500,000 under $1,000,000			
$100,000 under $200,000	Under $5,000,000	430,657	2,804	0.65
Under $100,000	$1,000,000 under $5,000,000			
$200,000 under $500,000	Under $10,000,000	113,977	2,516	2.21
Under $200,000	$5,000,000 under $10,000,000			
$500,000 under $1,000,000	Under $20,000,000	18,515	2,886	15.54
Under $500,000	$10,000,000 under $20,000,000			
$1,000,000 under $2,000,000	Under $30,000,000	5,776	2,931	50.74
Under $1,000,000	$20,000,000 under $30,000,000			
$2,000,000 under $5,000,000	Under $50,000,000	2,184	1,626	74.45
Under $2,000,000	$30,000,000 under $50,000,000			
$5,000,000 and over	Any amount	522	522	100.00
Under $5,000,000	$50,000,000 and over			
Forms 1040 only with Schedule F, but without Form 2555, Form 1116, or Schedule C		2,037,413	2,727	0.13
Under $20,000	Under $200,000	827,380	244	0.03
$20,000 under $50,000	Under $500,000	862,373	338	0.04
Under $20,000	$200,000 under $500,000			
$50,000 under $100,000	Under $1,000,000	257,595	280	0.11
Under $50,000	$500,000 under $1,000,000			
$100,000 under $200,000	Under $5,000,000	60,309	193	0.32
Under $100,000	$1,000,000 under $5,000,000			
$200,000 under $500,000	Under $10,000,00	23,230	255	1.10
Under $200,000	$5,000,000 under $10,000,000			
$500,000 under $1,000,000	Under $20,000,000	4,506	374	8.30
Under $500,000	$10,000,000 under $20,000,000			
$1,000,000 under $2,000,000	Under $30,000,000	1,327	481	36.25
Under $1,000,000	$20,000,000 under $30,000,000			
$2,000,000 under $5,000,000	Under $50,000,000	526	395	75.10
Under $2,000,000	$30,000,000 under $50,000,000			
$5,000,000 and over	Any amount	167	167	100.00
Under $5,000,000	$50,000,000 and over			
Forms 1040, 1040A and 1040EZ without a Form 2555, Form 1116, Schedule C or F		86,933,799	77,538	0.09
Under $20,000		50,758,547	17,879	0.04
$20,000 under $50,000		28,658,139	13,129	0.05
$50,000 under $100,000		6,477,008	10,342	0.16
$100,000 under $200,000		791,069	9,247	1.17
$200,000 under $500,000	Not applicable	205,073	9,222	4.50
$500,000 under $1,000,000		30,748	9,851	32.04
$1,000,000 under $2,000,000		9,085	4,600	50.63
$2,000,000 under $5,000,000		3,303	2,441	73.90
$5,000,000 and over		827	827	100.00

SOURCE: Internal Revenue Service, Statistics of Income, Individual Income Tax Returns, 1985.

- o weighting of the income components
- o indexing of stratum boundaries
- o representation of key items and subpopulations

We consider each of these problems in turn, offering as we do so a critical evaluation of the current sample design.

Inclusiveness of the Income Concept

The income totals on which the current stratification is based represent income subject to tax. Nontaxable portions reported on the tax return are excluded. These include tax-exempt interest income; nontaxable pensions, IRA distributions, annuities and rollovers; nontaxable social security benefits; and a variety of gains offset by losses. Prior to 1987, 60 percent of long-term capital gains were excluded as well.

The income concept also excludes amounts that are not reported on the individual tax return. Types of income not required to be reported include (among others) inheritances and bequests, welfare benefits, child support, veterans' benefits, workers' compensation, portions of scholarships, certain kinds of insurance benefits, and deferred compensation up to specified limits. Several of these items have the potential to represent large proportions of a filing unit's total income in any given year. However, the likelihood that these nonreported amounts constitute large proportions of any unit's income for several consecutive years is probably small except for lower income units.

Limiting the income concept to taxable components has certain drawbacks. First, definitions of what is or is not taxable can change; indeed, there have been significant revisions in recent years. Second, the exclusion of nontaxable components of total income can weaken the representation of subpopulations important for policy analysis--e.g., recipients of social security income. Third, taxable income may be less powerful than total income as a covariate of the full array of income and tax aggregates published by SOI. These aggregates include a number of nontaxable income items.

Treatment of Losses

As we noted earlier, several components of total income can be negative. The fact that large losses can offset large positive components makes special treatment of negative amounts desirable.

The separate summation of positive and negative amounts under the current sample design contributes to improved precision because returns with large offsetting amounts are sampled at a higher rate than if the stratification were based on the net amount. However, returns with negative totals exceeding their positive totals are classified on the basis of absolute values rather than placed in separate strata. Thus a return reporting a $40,000 total loss is included in the same sampling stratum as returns with $20,000 to $50,000 in total positive income.

Apart from reducing the total number of strata, the merits of combining negative and positive totals into the same stratum are not obvious. For one thing, large losses carry very different tax implications than large gains, so the precision of aggregate tax estimates is weakened by this tactic. Moreover, reported losses are more subject to manipulation by the taxpayer than most positive amounts, suggesting that negative totals are likely to be weak covariates of other income and tax items. Creating separate strata for returns with negative total income could improve the precision of both income and tax aggregates and at the same time ensure an adequate sample size for returns with negative total income--a potentially interesting subpopulation for tax policy research. If the number of returns with negative total income is too small to warrant establishing multiple strata, much of the same benefit in terms of variance reduction can be achieved by defining a single stratum for all returns with negative total income and post-stratifying on the magnitudes of losses.

Weighted Summation of Income Components

PAT and NAT are calculated as simple sums of their respective component items. While the ease of implementation and the intuitive interpretation of the resulting quantities have obvious merit, these totals are extremely sensitive to the values of individual components. For example, a return with a large capital gain from the sale of a residence but relatively small amounts from other income sources may be sampled at the same probability as returns with high amounts across a range of items. Income components with the highest variance and, therefore, the greatest impact on total income may not be strong covariates of other key items.

Table 2 reports the correlations among selected income items, including total income subject to tax, within two subsamples of the tax year 1979 SOI sample. The correlations above the diagonal were calculated from returns designated for the Continuous Work History Sample (CWHS), which represented a five in 10,000 sample of tax returns filed in 1980. The correlations below the diagonal were calculated from the very high income returns selected with certainty into the SOI sample. Most of these returns met either of the following selection criteria: (1) AGI or a component amount of $500,000 or more, or (2) total receipts of $5,000,000 from a business and/or farm. The handful of returns with membership in both the CWHS and high income subsamples were included with the latter in calculating the correlations reported in Table 2.

With few exceptions the correlations among the individual income items are very small. Most of the items are correlated much more highly with total income than with other components. Among the CWHS returns, salaries and wages are by far the strongest correlate of

Table 2--Correlations among Selected Income Items, 1979

	TOT	SAL	INT	DIV	CAP	SCE	PEN	UNC
TOT	–	.637	.302	.283	.490	.342	.027	-.016
SAL	.118	–	.015	.036	.012	-.089	-.041	-.004
INT	.206	-.018	–	.248	.065	.094	.074	-.017
DIV	.268	-.054	.156	–	.077	-.049	.033	-.011
CAP	.711	-.033	.136	.072	–	-.070	.004	-.003
SCE	.231	-.063	-.050	-.076	-.053	–	.083	-.008
PEN	.009	-.008	.007	.025	-.003	.012	–	-.011
UNC	-.004	-.004	-.002	-.003	-.001	.000	-.001	–

SOURCE: 1979 SOI Individual Sample.
NOTE: Correlations above the diagonal were calculated from the CWHS subsample (N=45,840); correlations below the diagonal were calculated from the high income subsample (N=19,046). The individual items are defined below:

TOT	Total income subject to tax
SAL	Salaries and wages
INT	Interest received
DIV	Dividends in AGI
CAP	Combined net capital gain or loss
SCE	Schedule E net income or loss
PEN	Taxable portion of pensions and annuities
UNC	Unemployment compensation, total

total income. Among the high income returns, however, salaries and wages are only weakly correlated with total income while capital gains are strongly correlated with total income. Pensions and unemployment compensation are not correlated with total income in either subsample--a fact which we can attribute to their infrequency and to the fact that they are more likely to be present on relatively low income than high income returns.

As an alternative to the simple sum of the income components, a weighted sum would provide a better stratifier, theoretically. Components that are relatively strongly associated with the full range of income and tax items would receive greater weight than components that are only weakly associated with these items. The weights would be derived empirically from the covariances between the set of items for which aggregate statistics are to be produced and the prospective components of the stratifier.

The weighting need not be the same across all returns. Table 2 provides evidence of substantial differences in the inter-item correlations among very high income returns versus all other returns. This suggests that there is a potential to improve the stratification even further by applying differential weighting schemes to the income components among different classes of returns.

An obvious disadvantage of the weighted summation is the possible need to re-estimate the weights periodically, should the underlying relationships prove to be unstable over time. Even without such re-estimation, however, the relative advantage of weighted versus unweighted summation would likely be maintained.

Indexing of Stratum Boundaries

Under the current sample design the boundaries between strata are not indexed for inflation. With one recent exception, the boundaries have been fixed in nominal dollars since 1982. Income growth, both real and inflationary, shifts the distribution of the population by stratum. Sample size constraints make it necessary to adjust the sampling rates by stratum each year. Growth in the strata sampled at 100 percent necessitate reductions in the sampling rates in other strata. Over time, the sample drifts increasingly from the original design.

The sample drift and the need for annual adjustment to the sampling rates could be reduced by indexing the stratum boundaries to constant dollars. Indexing poses a number of problems, however. These include the initial research required to identify a suitable index or multiple indexes applicable to different sets of strata, the need to project index values each year (because the sampling rates must be specified before the index values for the year can be known), and the fact that the sample drift to which some users may already be accustomed will be altered. To the extent that users are oblivious to the sample drift that occurs in the absence of indexing, this last problem is less important.

Representation of Key Items and Subpopulations

Policy analysis of tax issues requires an ability to assess the impact of a tax law or proposal upon certain well-defined subpopulations--e.g., retired persons, the poor, the wealthy, and middle income families. However, it is also important to identify the specific segments of the population that are most affected by a given tax law or proposal, and these groups are defined only by the presence of particular line items and filing characteristics.

The current sample design includes a small number of specialized strata that were inserted to meet the needs for data on specific tax schedules or forms, and on filers with particular characteristics. In addition, the current stratification provides large samples (or 100 percent representation where the populations are small) of high income returns. However, the sample sizes for low income returns and possibly returns filed by the elderly may be too small to adequately serve policy analysis needs. Similarly, line items that are used most commonly by lower income or elderly taxpayers may be too rare in the sample to support analysis of the impact of change.

That the current design does not explicitly address these needs may be attributable in part to the difficulty of defining such needs and of creating a sample design that will take account of them. Determining what subgroups or line items merit explicit attention in the sample design requires anticipating the areas of legislative activity in the tax arena several years hence. Improving the representation of different demographic subgroups is complicated by the fact that the taxpayer's age does not appear on the tax return and cannot be used for selection. Finally, the budgetary constraints on sample size imply that the low income sample can be increased significantly only by reducing the sample of higher income returns. Changing the sample allocation without revising the stratifier to improve its efficiency would exact a price in terms of the variances of the estimated income and tax aggregates.

RESEARCH DESIGN

The research agenda that has been developed in an effort to define a new income stratifier for the SOI individual sample will include the following elements:

- o examination of returns reporting losses on one or more major income components
- o estimation of weights relating the prospective components of a stratifier to the income and tax items for which aggregate estimates are required
- o investigation of the representation of key sub-populations
- o simulation of alternative sample designs to estimate item variances and subpopulation sizes

These elements of the research design are discussed below. First, however, we discuss the data sources that will support the empirical research.

Description of Data Sources

The data that will be used for research to develop the new stratifier include the SOI "complete report" files for the 1984 and 1985 tax years, plus the supplementary sample files for those years. The complete report files, which include about 83,000 records in even-numbered tax years and about 121,000 records in odd years, contain data edited specifically for statistical uses and are the basis of the annual SOI publications on individual income tax returns.

The supplementary files include more limited, unedited data for three samples. The Level 3 sample consists of about 300,000 records, including the complete report sample records plus additional records selected under the same overall design but with higher sampling rates. The Level 4 sample is drawn by applying the Level 3 selection criteria to the secondary social security number (SSN) on joint returns. (The role of the SSN in sample selection is described in a later section.) The Level 5 sample includes all returns with primary SSNs that were selected into any level in a prior year (beginning 1982) but not the current year. Together Levels 4 and 5 add about 260,000 returns to the 1984 and 320,000 returns to the 1985 sample.

Other pertinent features of these data are discussed in the sections that follow.

Returns with Losses

There are two issues with respect to the treatment of returns with losses. The first issue is how to count

losses in assigning returns to income strata. The second issue is whether and how to separately stratify returns with net or large losses.

The treatment of losses becomes important only when there is a potential impact upon stratum assignment. The probability of such an impact rises as the losses grow large relative to the income class boundaries, but even a small loss can affect stratum assignment if the positive income sum is only marginally greater than the nearest stratum boundary. The potential significance of all such cases must be weighed in determining how to count reported losses.

Distributional information on both the absolute and relative magnitudes of losses reported on tax returns are crucial to the determination of an appropriate strategy for dealing with losses. To satisfy this need we will begin by constructing a distribution of negative sums by positive sums for a representative sample of 1985 returns. What we learn about the frequency and magnitudes of losses will influence how we approach the development of weights for the individual income components (see below), as we may wish to apply differential weights to positive and negative amounts. We will recreate the tabulation with weighted sums once we have developed preliminary weights.

The decision on how to stratify returns with large losses will depend on the frequency of such returns in the population, their distribution by size, and a judgment as to whether the goals of stratification include insuring a minimum sample size of returns with very large losses. The latter may be of interest to Treasury Department analysts. If not, then we may be able to achieve the single goal of improving the precision of estimated income and tax aggregates by creating a single stratum for returns with large losses, however such losses come to be defined, and post-stratifying.

Estimation of Component Weights

The problem of estimating weights for the individual components of the income stratifier can be generalized to encompass the determination of what components should be included in the income concept. The research strategy is as follows. Given the set of prospective components and the set of income and tax items for which aggregate estimates are required, we wish to determine what linear combination of the components will maximize the covariance with the latter set of items. We will work with a subset of the 200 items, relying upon expert opinion (namely SOI staff and the Division's principal clients) to prioritize the items.

In estimating weights we will exclude the specialized strata, and we will estimate separate models within the business, farm and other classes of returns, as their separate stratification will be maintained under the revised sample design. In addition, we may divide each class of returns into two or three broad groups based on some type of total positive income criterion, so that we can determine to what extent the relative importance of different income components varies by broad income level. For example, based on the correlations reported in Table 2, we would expect to find that salaries and wages receive a much greater weight among lower income returns than among higher income returns.

Continuity of Stratum Membership over Time

One of the other elements of the individual sample redesign is the introduction of a large and representative panel, to be followed for a period of several years. While the initial panel will be drawn from returns selected under the current stratification, future panels will reflect the revised stratification. Consequently, there is a need to consider longitudinal aspects of the stratification.

An individual's total income in any year can be interpreted as the sum of a "true" income level, consistent with a long term trend including past and future experience, and a disturbance, reflecting short term influences. Averaged over similar taxpayers over a sufficiently long period of time, the disturbances net to zero. In this view the most appropriate stratifier for a sample that will be followed over time is true income. Stratifying instead on the total income observed in the year of selection will result in a less efficient and perhaps biased sample. In particular, if the probability of selection increases with total income, individuals with positive disturbances will have a greater chance of selection than individuals with the same true income but negative disturbances. Over time the pattern of change in these individuals' total income will deviate from the pattern that would be observed among all individuals.

One way to assess the extent to which a stratifier captures the taxpayer's longer term income status is to examine movement among strata over time. The data that we are using will include 1985 returns for most of the edited 1984 sample, including many individuals who shifted from primary to secondary filer. We will examine transitions in filing status and movement among strata between the 1984 and 1985 tax years.

Even in the absence of the panel element in the individual sample redesign, information on the longitudinal dynamics of income components and prospective income stratifiers could be valuable to the development of a good stratifier. Differential weighting of the income components, if derived in the manner discussed previously, will yield income totals similar in concept to true income. Longitudinal information can be used to refine the weighting. Earlier we cited the potential volatility of some income components as a drawback to calculating the income stratifier as a simple sum of its components. The variability of a particular component over time is a problem if that variability is not associated with change in other income and tax items. With the linked 1984 and 1985 data we will be able to expand our investigation of component weighting to include an examination of the between-year variation of individual income components. As a preliminary step, this examination of between-year variation may be helpful in screening the component income items. It may also prove useful in determining how to deal with reported losses.

The stability of stratum membership is relevant to the sample design in another way. To understand this, one must know how the SOI sample is selected. Within each stratum (but excluding the CWHS subsample) the selection of returns is based upon a transformation of the primary taxpayer's SSN. After truncation the transformation yields a pseudo random number which is compared to a target number for that return's stratum. Returns with transforms below the target number are selected into the sample.

The transformation algorithm is held constant from year to year, so that a given SSN always produces the same transform. Consequently, a particular SSN, once selected, will continue to be selected as long as the taxpayer's return falls into a stratum with the same or higher sampling rate. A taxpayer who drops into a lower stratum will face a reduced probability of selection. Examination of the sampling rates in Table 1 shows that the chances of being dropped from the sample are much greater for some cross-stratum moves than others. The largest reductions from single stratum crossings between 1984 and 1985 would have occurred to taxpayers filing without Schedules C or F, who dropped below the $500,000 or $100,000 income levels. Taxpayers making these transitions had only a one in eight chance of being retained in the SOI sample.

Evidence from a few years ago suggests that about one third of the SOI sample turns over between consecutive years. Individuals who drop out of the filing population or file as secondary taxpayers account for some of the turnover. While detailed information on sample turnover is unavailable, however, we suspect that changes in stratum membership account for a significant proportion of the lost returns. For example, we have determined that 38 percent of the taxpayers who met the selection criteria for the 100 percent strata in 1979 (these were primarily very high income returns; see above) did not meet those criteria in 1980.

Representation of Key Subpopulations

It is possible to link demographic data from Social Security Administration files to the returns represented on the 1984 and 1985 SOI files. Once this is done we will be able to determine the age composition of the SOI sample under the current design and under prospective new designs (see the final section). We will also be able to explore tactics for increasing the representation of individual age groups under the revised design. Such tactics must make use of items that are available at the point of sample selection. It may turn out that increasing the size of the lowest income stratum is the most efficient way to enhance the representation of several demographic groups whose sample sizes are currently too small to support policy analysis.

Tax policy modeling at the micro level frequently involves the use of values reported in individual fields on the tax return to estimate the impact of changes in the tax treatment of those items. We will examine the representation of individual line items in the SOI sample currently and under prospective revisions. As an initial step we will estimate both the sample and population frequency of each of the 200 income and tax items for which aggregate estimates are reported in SOI publications, and we will submit this report to the SOI Division and to the principal SOI clients for review. In view of the significance of recent tax changes, we may opt to produce these tabulations from preliminary 1987 data.

The SOI Division or its clients may determine that the sample sizes of certain items are too small for reliable estimation of policy impacts. If so, we will need to consider tactics for improving the representation of these items. Increasing the sample frequency of rare items, however, is likely to require more than simply increasing the sample size of an entire stratum. If the number of items is small or if the items tend to occur on the same returns, the creation of a specialized stratum with a relatively high sampling rate may be a viable option that will not unduly complicate the sample design.

Another option that we will investigate would involve separating the lower income classes into additional strata based on the complexity of the tax return. The more items that are present on the return, the greater the use that can be made of that return in policy modeling. A return with only wage and salary income does not enhance the modeling capabilities of the data base except with respect to the basic tax rates. Oversampling returns with large numbers of reported items is particularly sensible when the potential areas of policy activity are broad and, to a significant degree, unknown.

Simulations

The analyses described thus far will provide little information on the actual performance of a new sample design with respect to the variances of the tens of income and tax aggregates that are of interest, or to the resulting sample sizes of key subpopulations. To obtain such critical information we must develop a simulated population of tax returns, which we can then stratify and sample in alternative ways to produce the statistics that we require.

The simulated population will consist of a suitably weighted sample much larger in size than the target of 93,000 for the revised sample design. The Level 3 SOI sample for 1984 or 1985 might be sufficient for our needs. This sample, discussed above, includes about 300,000 returns and can be weighted to national totals. There are about that many additional returns available each year in the Level 4 and 5 samples, but these returns are not weighted, and appropriate weights for the entire set of returns cannot be constructed without an undue investment of resources. However, some of the Level 4 and 5 returns may be of use in expanding selected strata in ways that do allow proper weighting.

The simulated population will provide a data base from which we can estimate the stratum variances needed to develop an efficient sample allocation scheme. The calculation of these estimated variances will be the first step in the evaluation of a prospective stratifier.

The need to satisfy a number of constraints in developing the sample design prevents the straightforward derivation of an optimal allocation of sample sizes among the strata. These constraints will include:

- o fixed sampling rates of 100 percent for selected strata
- o a fixed minimum rate of .02 percent due to the CWHS subsample
- o probable constraints on the minimum size and distribution of the Schedules C and F returns
- o possible constraints on the minimum sample sizes of key subpopulations

In developing alternative sample designs we will proceed, generally, by first determining the initial constraints, then estimating a sample allocation, and then drawing a sample to determine whether the additional constraints have been satisfied. If further adjustment of the sample allocation is required, we will make such adjustments, draw a new sample, and then estimate the item variances as well as the subpopulation sizes.

ACKNOWLEDGMENTS

This research is being performed under contract to the SOI Division of the IRS. I am grateful to Fritz Scheuren and to the members of the Individual SOI Redesign Planning Team for providing important background information and stimulating suggestions, and to Roderick Little and Donald Rubin for their significant contributions to the research design. I also wish to thank my Mathematica Policy Research colleagues Bob Cohen and Gary Swearingen for programming assistance, and Patrice Turner and Lena Cunningham for manuscript preparation.