STATISTICS OF INCOME: AN OVERVIEW

Robert A. Wilson and John DiPaolo, Internal Revenue Service

In December 1980, the Statistics Division of the Internal Revenue Service prepared for consideration its first multi-year operating plan, in part to meet the directive of the then Office of Federal Statistical Policy and Standards (OFSPS) and in part to meet the requirements of the first IRS "strategic plan".[1] These new reporting requirements now give users more of an opportunity to review the long-range plans for the Internal Revenue Service Statistics of Income (SOI) program than was provided formerly.

This paper is based on material included in the introduction to the long-range plan and reviews some of the major procedural and methodological strategies being considered for the future. The presentation begins with an introduction to the SOI program as background, an explanation of the general concerns that have been raised about resource needs relative to the program, and a summary of how SOI data are now processed. This is followed by examining several of the processing innovations which will be researched and evaluated for possible implementation during the planning period as a means of increasing productivity.

THE STATISTICS OF INCOME PROGRAM

The Internal Revenue Service, in addition to its primary mission of enforcing compliance with the Federal tax laws, is also charged with the responsibility of publishing statistics on the operation of these tax laws. The data, based on tax returns, are published in a series of reports called Statistics of Income.

This series came into being soon after adoption of the Sixteenth Amendment to the Constitution and the subsequent enactment of the first modern U.S. income tax law, the Revenue Act of 1916. The Act specifically called for the annual publication of statistics. The wording contained in the 1916 Act has been repeated, with practically no change, in each major rewrite of the Internal Revenue Code since that time. It is currently contained in the 1954 Code, which is the basis for the current tax law.

The SOI reports from the very beginning (1916) have been used extensively for tax research and for estimating revenue, especially by officials in the Department of the Treasury. At the start, the reports were geared almost entirely to meeting these needs. With the growth of research groups both within and outside of the Federal Government and with the increased needs of tax planners and revenue estimators, new types of data soon were also required. At the same time, the tax returns were expanded to reflect the growing number of new provisions of the law, thus providing a ready source with which to meet these needs.

By the close of World War II, most of the population was subject to the income tax. At about the same time, the economies of using existing administrative files as the source of data on a wide variety of statistics had become more and more apparent. While the tax definitions of data items presented some obstacles, the obstacles were far outweighed by the likelihood that taxpayers' response tended to be more accurate than their response to special surveys. Moreover, with experience, users learned how to adjust for these definitions to meet their own particular needs.

The upshot of all these developments was an SOI increasingly different in its orientation from the early SOI. Several multi-purpose reports replaced the single tax-oriented report. While tax data continued to be included (all the more so as the tax law expanded both in scope and in complexity), the emphasis changed to more general purpose statistics geared to meeting the needs of economists and financial analysts.

The main emphasis of the annual statistics has always been individual and corporation income tax data. Other subjects based on other types of returns for which data have been tabulated either annually or periodically have been partnerships, estates and gifts, fiduciaries, farmers' cooperatives, foundations and other tax exempt organizations, and employee plans. Schedules attached to some of the returns become the subject of their own SOI reports. The sole proprietorship schedules were a relatively early source of statistics, which together with data from partnership returns, shed light on an important part of the economy not covered anywhere else to any appreciable extent.

Another development in the growth of SOI was the increasing tendency for new revisions to the tax law to require separate reports to Congress by Treasury's Office of Tax Analysis (OTA). These reports required statistics on such topics as individuals with high income who were nontaxable, the operation of the jobs credit provisions, Domestic International Sales Corporations (DISC's), international boycott participation, taxation of corporate income from U.S. possessions, and income of citizens working abroad.

Organizational Relationships

The Statistics Division in Washington is part of the IRS Office of Planning and Research. This office plays a leading role in developing taxpayer compliance studies and quality control systems, conducting new systems feasibility studies, and in identifying administrative problems in adapting to new law changes. The Statistics Division is responsible not only for SOI, but also for supplying IRS long-range workload projections and for conducting special statistical studies for the Service and supplying advice on sample designs for use in helping other organizations in IRS conduct studies of their own.

In connection with SOI, a staff of statisticians and economists works closely with users to determine the content of each program and publication, to design the samples used, and to develop field procedures. Complications arise from the fact that the processing is decentralized in twelve different locations throughout the country (see figure 1); hence there is a need for a strong coordinating role by the Statistics Division, including adequate quality controls to assure uniform and accurate processing.

The SOI program has the following basic character. Returns filed with the ten service centers are processed for administrative purposes to determine the correct tax liability. During processing, the returns are entered on tape for eventual posting to the IRS Master File. It is when the return records are on tape that they are selected for SOI. After the returns are selected, they are subjected to additional editing for SOI by specially trained technicians. The data thus extracted from the sample returns are entered on tape and tested for consistency. Any errors detected are then resolved to produce a final data file which is used to prepare SOI tabulations.

SOI Users

Information obtained from the SOI program is used extensively throughout the Federal Government for a variety of purposes. Besides OTA and the Joint Committee on Taxation, the third major Federal user of SOI is the Bureau of Economic Analysis (BEA) in the Department of Commerce. Data on corporations in the National Income and Products Accounts [2] are benchmarked to the amounts reported on corporation income tax returns which are then adjusted for conceptual differences and extrapolated based on more fragmentary data from other sources. Returns of unincorporated businesses, i.e., for sole proprietorships and partnerships, are also used for the national accounts; they constitute the only complete and reliable source of financial statistics for this segment of the economy. Investment income from individual income tax returns is also used in the national accounts.

In prior years the detailed planning for an SOI year began with user meetings which were held during the spring of the tax year under consideration. These meetings were attended primarily by representatives from OFSPS, OTA, Joint Committee on Taxation, and BEA; some of the other agencies that also participated included the Social Security Administration (SSA), Bureau of the Census, Federal Trade Commission, Department of Agriculture, and Small Business Administration.

The format for these meetings consisted of presenting the users with a marked-up copy of the tax forms or return schedules showing which items were proposed for inclusion in SOI for that year. These proposals were based on the frequency or content of recent prior-year programs that were reflected in previous plans; informal discussions held earlier at lower management and technician levels; known or anticipated law changes for which data would likely be needed; and, of course, the extent of available statistical resources. Often, because of lead-time constraints, only limited changes to the proposed program content were possible.

NEW PROGRAM CONTENT STRATEGIES

The basic assumption used in developing the present multi-year strategic plan was that the demand for statistical data was likely to increase in the 1980's and that resource constraints on Government statistical programs would probably continue. To this end, the Statistics Division recently reevaluated the size of each of the SOI samples and presented a new plan to its major users.

The resulting sample size reductions are to be coupled with improved methods of weighting the data. The introduction of post-stratification in all SOI programs is being examined as a possible means for maintaining reliability in the face of new sample size reductions. These reductions are to be accomplished by basing the estimates on subsamples of the former full sample sizes of the late 1970's; the larger samples will continue to be designated, but their use will be confined, for the most part, to improving the weights for the subsample. The larger samples will also be available for reimbursable projects (see figure 2).

Another strategy under examination is the separation of program content into "core" and "other". The core programs would generally be stable, from year to year, and would consist of the basic elements of each program which change only occasionally, when the law or tax forms change. The rest of a program would continue to vary from year to year to meet the changing needs of tax policymakers.

The core program for individual income tax returns would consist of the various sources of income, personal exemptions and deductions, income tax computation, tax credits, and tax payments. The "other" program could consist of studies of the minimum or maximum tax computation schedules, sales of capital assets by type and computations of various tax credits, to cite some examples. In the case of corporations, the core program might consist of the income statement, balance sheet, income tax computation, tax credits, tax payments, and distributions to stockholders. Thus the "other" category could consist of computations of the investment, foreign tax, targeted jobs and work incentive credits and of the mimimum tax. Anything else could either be a Treasury Special Project, or a reimbursable project under this proposal.

Statistics for the core program would be produced in such a way that the entire computer system would not have to be redesigned to facilitate its processing each year. To be consistent with this, more of the statistical table outlines would also remain the same from year to year. Manual and computer processing would thereby remain constant with resultant economies.

When computer programs could not be simply updated, because of the necessary changes in the SOI program content, increased use of generalized systems would be substituted, thereby still achieving a net saving. Only a limited amount of data from the non-core program would be published and only in summarized form; the extent to which special OTA items are used further, such as in the SOI reports, would be dependent on OTA's needs.

DATA ABSTRACTION FROM RETURNS

For most SOI programs, up until now, Master File data have been used sparingly because of their limitations.[3] Until recently, the primary use made of Master File data for SOI had been in identifying returns for the samples used and for advance or early tabulations to meet special requests.

Beginning with Tax Year 1981 or 1982, manual editing or data abstraction from returns for statistics using a specialized abstract sheet will become economically obsolete for many programs. Instead, return data for the SOI sample will be obtained from the Master File system. When possible, adjustments to overcome shortcomings in the Master File data will be introduced through computerized routines. This method will be gradually extended to all SOI programs.

Every five years, a more comprehensive manual statistical edit, often involving many more items than are available from the Master File system, might take place for the SOI unincorporated business programs, possibly using an abstract sheet. This special editing would coincide with the Agricultural and Economic Censuses planned for 1982, 1987, etc. Special requests for data may be accommodated in a like manner. For example, the Department of Agriculture has expressed interest in obtaining tax return statistics on farming activities in addition to information that would normally be provided as part of SOI for use in connection with the Agricultural Census.

Since the cost to Agriculture of obtaining the required information through conventional survey methods is prohibitive, it may be possible in the future to increase the farm portion of the SOI sample to obtain this information for them on a reimbursable basis. In the interim years, changes in program requirements would be kept to a minimum so that all programming and manual instructions may be held constant to the maximum extent. This would facilitate meeting completion dates for major functions in each program, thereby speeding delivery time of the final product to users while conserving both professional and clerical resources.

For those SOI items which are not key-entered to the Master File tapes during revenue processing, an abbreviated abstract sheet may be required. The size of the sheet, however, will be kept to a minimum, providing perhaps for only those items that are to be manually abstracted. Under this approach, data from the Master File system would be transferred directly to an SOI tape for later consolidation with the manually-edited items.

Current thinking is to base some SOI programs, namely individuals, sole proprietorships, partnerships, and fiduciaries, almost entirely on Master File information. These data may be augmented each year, to a limited extent, by additional data that are manually edited for statistical purposes and that are not available through the Master File, although how this might be done is still being explored. For the annual individual income tax return statistics program, the number of Master File items available will be far more numerous and comprehensive than for the unincorporated business and fiduciary programs. For corporations, the relatively few data elements for SOI that are transcribed for revenue processing are currently under study in order to determine the extent to which they can be utilized for SOI; their use may be possible at least for smaller corporations in the SOI sample.

The current explorations will also determine whether there are some relatively inexpensive changes that can be introduced into the administrative processing system which would facilitate statistical use of Master File data. These might include the processing of limited additional data elements now not required for administrative processing.

To the extent such steps can be accommodated at this earlier stage in return processing, added costs at later stages, i.e., during statistical processing, may be avoided. Items still not used in administrative processing, or for which adjustments during administrative processing are inconsistent with their use for statistics, may be obtained as in the past by manually abstracting the data in an off-line statistical processing operation. In some cases, this may be facilitated by use of specially designed, smaller, samples for this purpose; presently a general-purpose sample is used for all statistics from a given return form.

COMPUTERIZED EDITING, ERROR DETECTION AND CORRECTION

Integration of the two sets of data, from the Master File system and from the statistical processing system, will be facilitated by a computerized error resolution system which would increase the role of the computer either in editing certain data which were manually edited in the past or in estimating data missing from the returns as filed. To the extent that this can be accomplished, in part with the aid of prior-year "perfected" statistical data for the same taxpayers, a more economical substitute for former procedures may be achieved.

For some programs, more of the computerized testing of each record for internal consistency testing and error resolution associated with this testing will take place concurrently with editing to shorten the feedback cycle to editors, verifiers, and data transcribers and to enable the correction of errors while the tax return is still available.

Much of the return editing will be computerized as part of this operation, thus replacing to a varying extent, the former manual operation. While past studies point to significant problems in any extensive use of Master File data without some form of statistical verifi- cation, the plan now under development calls for flushing out discrepancies, insofar as possible, using the computer to identify returns with computations "out of balance" or with other problems. Only the returns that fail this preliminary screening would be manually edited. This approach assumes some redefinitions of data items now manually edited because certain adjustments now made in manual editing might not be identifiable by computer. The extent of these redefinitions will depend on the SOI program under consideration.

At the same time, an automated approach is contemplated that will deal with schedules and items missing from the return. For example, a significant number of partnership returns are filed with balance sheet or other data missing; research is therefore needed to develop a methodology for the imputation of this missing information.[4] The Statistics Division is actively seeking outside funding for this purpose. For other returns, identification of missing schedules and items early in processing will permit followup to obtain various missing data in time to prevent delays later on in processing.[5]

The new methodology would contribute to a lower cost of controlling overall data quality because of the reduced error rates following the initial institution of more timely feedback of error conditions to the originators. Longitudinal characteristics of the sample would be used to advantage in consistency testing. Selected ratios based on tax return data would also be computed for comparison to the prior-year's ratios. For the business and corporation programs, industry codes would be systematically compared to prior-year codes to detect gross errors. Many of the errors would then be corrected by computer, while errors of a more complex nature would be read out for resolution by professional subject-matter staff members in the Statistics Division.

Finally, the IRS is currently engaged in a study to evaluate a new overall system for handling key entry and error resolution. The present system involves many hours of complicated separation of printed registers, and the association of registers with the related returns or other input documents. Error resolution clerks must then manually correct the register which is then batched and controlled for key entry. The use of on-line systems are now under study. These would utilize direct access to documents in error through a terminal that is connected to a minicomputer, permitting the corrections to be made without intermediate processing. We look for this approach to have an important long-run beneficial impact on the SOI program.

The success of new approaches to or substitutes for the present statistical editing process and of the expanded use of Master File data will be largely dependent on the adequacy of a quality control system. Presently, the quality control system that is used in statistical processing is concerned mainly with the effectiveness of the data abstracting or editing operation. Its major limitation is lack of timeliness for corrective action. The "system of the '80's" will check, not only on the manual editing (for those programs for which manual statistical editing is still applicable), but also on the processing at each subsequent stage, so that it will be possible to identify on a more timely basis the exact stage at which changes to the "original" data are made for any given return. The appropriateness of the changes made can then be more adequately assessed. As a byproduct, additional measures of nonsampling error will become available.

Industry Coding

Currently, most of the business and corporation tax returns are industry coded by the taxpayer using the numbered groupings that appear in the return form instructions and that are based, for the most part, on the Standard Industrial Classification. (For sole proprietorship schedules, the IRS attempts to code the return based on the taxpayer's description in the absence of a perceived need for a self-coding requirement.) An independent statistical coding operation is now included for returns selected for the SOI samples and involves, in general, consistency of the reported code with other information from the return itself (including the source of the receipts shown on the return and the business' narrative description of its principal business industrial activity and product) or from reference books. It is estimated based on the results of this independent coding that up to one-third of the self-coded entries may be in error. Therefore, the taxpayer-reported codes which are transcribed in revenue processing are not acceptable for most statistical purposes. On the other hand, economies may be realized if perfected codes can be obtained elsewhere in Government, either annually or periodically. These codes could be used each year in place of those reported by the taxpayer. To accomplish this, legal and practical problems would first need to be overcome. The former involve confidentiality rules affecting IRS and other agencies; the latter involves differences in the statistical reporting unit among agencies which could limit the appropriateness of any interagency use of a given code for a given business. [6]

The longitudinal aspects of the basic business samples might permit increased utilization of the SOI industry code from the prior year.[7] The SOI industry code previously obtained would be used; then, if the taxpayer's self-reported present and prior-year's code were the same, the prior-year SOI code would be used again without further research. On the other hand, if there were a difference in the taxpayer's industry code from year to year, the return would be examined to determine if there appeared to have been a real change in business activity. Among

other things, this type of two-year comparison would result in more stable estimates of industry from one year to the next at less cost.[8]

EXPANDING THE SOI DATA BASE

If SOI is to serve tax policymakers in a more responsive manner and on broader issues, it will be necessary to build a data base from as many sources as possible. With this in mind, the Division is now establishing exchange agreements with other agencies with regard to information furnished to them by the Internal Revenue Service under provisions of Internal Revenue Code section 6103, as amended by the Tax Reform Act of 1976 (which limits access to return records to specified governmental agencies for specified purposes). The new agreements will provide that the IRS, on request, will be entitled to receive back a copy of the information furnished which will also include any perfection, modifications, or enhancements, or the addition of any other information prepared by the other agency for inclusion in, or for use with, the IRS-supplied data (to the extent possible, given the confidentiality rules of the other agencies).

The larger data base made possible by the inclusion of data from other agencies would make the Division more responsive to the research needs of other activities within the IRS and within the Treasury as a whole. Combined uses of SOI and the IRS Taxpayer Compliance Measurement Program are contemplated, for example.[9] As another illustration, working with SSA and the National Cancer Institute, Statistics Division would be able to provide mortality and morbidity data within demographic subgroups by an individual's occupation and industry.

Considerable research is, of course, necessary to develop or perfect methods of overcoming the many known difficulties that would be encountered in trying to expand the data base. For example, techniques would have to be developed for linking employer, taxpaying entity, establishment, pension plan, payroll entity, and employee. Such linkages would encompass all types of employers, including corporations, sole proprietorships, and partnerships.

Long-range plans might require the addition of an individual taxpayer's sex and age to the Master File system, along with an occupation code. Age and sex could be obtained from SSA files. Inclusion of age would permit a study of the relationships between income and age, and measurement of income differences between individuals with income from different kinds of retirement plans and individuals with no income from formal retirement plans. The existing SOI sample design results in an oversampling of individuals at the peak of their income-producing years. Including age in the Master File would permit stratification of the SOI sample to yield better measures of income for both younger and older taxpayers.

REDUCING SOI PUBLICATIONS

The SOI reports for the 1980's will be streamlined in that they will emphasize the presentations that change but little each year. The more dynamic presentations highlighting data on detailed computations from the tax return may be presented only in short summary tables. Besides the basic SOI reports, vehicles for releasing statistics could be news releases or special supplemental SOI reports, such as those already used to shed light on the foreign tax credit and on sales of capital assets, for example.

The 1980's are expected to witness a continuation of the trend already well underway, namely, direct employment by SOI users of the microdata records on computer tape. While disclosure rules effectively limit the extent to which this can now occur, it is expected that public use files containing microdata in a form not inconsistent with the current IRS disclosure provisions will be developed in the next few years and that their use will no longer be restricted to Treasury and to those other users now already authorized under the law to receive these data. Much more research needs to be done in this area, and much better documentation on the content of the SOI tape files as they already stand will be required, too. This initial investment can be expected to be costly in time and resources.

Other, perhaps short-run, solutions to more timely release of the SOI complete report statistics will include elimination of the preliminary reports long associated with the major SOI programs. For many years now, about half of the preliminary reports have been based on early cutoffs of the samples. However, for corporations, in order to produce meaningful estimates based on an early cutoff, an elaborate system had to be developed in order to estimate data for returns of many of the larger corporations. Elimination of the processing steps unique to the release of preliminary data, such as in the case of corporations, can lead to concentrated efforts, resource-wise, to develop a single system for each program in order to perfect data for the complete reports on a timelier basis.[10]

This curtailment will present a void, however. A publication vehicle was recently developed in the SOI Bulletin; the Bulletin is a quarterly report, that began with the summer issue which was released in July 1981. In the future, this report will include an advance release of selected tables from forthcoming SOI complete reports, as a partial substitute for the former preliminary SOI reports. The Bulletin will also include, among other subjects, tabular summaries of early data based on the Master File system. These Master File data are now produced routinely each month based on individual income tax returns for use by IRS, OTA and the Joint Committee on Taxation. More fragmentary data from the Master File are available annually for corporations and tax-exempt organizations which may also be included in the Bulletin.

CONCLUDING COMMENTS

Streamlining the SOI programs is not confined to cutting the size of samples, programs, and publications. Methodological and processing changes have to keep pace or even lead the way. The proposals to introduce concurrent computerized consistency testing of the data while SOI returns are still accessible, and to make more use of data for other years for the same taxpayer in perfecting return data for the current year, have already been mentioned. Other innovations, now well along in development, include use of generalized systems and of electronic composition as a substitute for typesetting tables to be published. Neither of these steps is a true innovation; rather, each is an example of steps that would have been introduced earlier, had resources been available with which to conduct the needed research. In fact, most statistical agencies have long since made use of them in their own programs.

A Generalized Tabulating System (GTS), initially developed by the Census Bureau, is now already in use in developing the tables for some SOI projects. Attention will now need to be focused on developing a generalized system applicable to "front-end" processing of the return data themselves, including the consistency testing and any automatic error resolution. Complete tape-to-tape electronic composition is soon to be phased in for use in all SOI reports.

Savings realized from economies due to reduced samples and more efficient methods of data processing will enable the Statistics Division to meet the needs for more statistical data expected in the '80's, and to release the regular SOI reports and studies on a more timely basis. They should also enable the Division to devote increased resources to new areas of research and to satisfy the needs of its major users.

ACKNOWLEDGMENTS

The authors wish to thank Ross Summers and Ralph Bristol for reviewing the manuscript copy of this paper, Wendy Alvey and Beth Kilss for their help in presenting this paper at the Annual American Statistical Association meetings, and Terry Smith for preparing the illustrations. Thanks are also due to Denise Herbert who typed the several drafts of this paper and to Bettye Jamerson who edited the manuscript.

NOTES AND REFERENCES

[1] The OFSPS requirements were stated in the Statistical Reporter, May 1980; the Internal Revenue Service requirements were defined together with the results in the report entitled Strategic Plan for the IRS, December 1980.

[2] See the Current Business Statistics published monthly in the Survey of Current Business, Bureau of Economic Analysis, U.S. Department of Commerce,

[3] Data are "perfected" for administrative processing only to the extent they have a direct bearing on the ultimate computation and verification of tax. However, not all of the procedures are consistent with statistical needs.

[4] Internal Revenue Service follows up through correspondence with the taxpayer on only selected schedules found missing during administrative processing of the returns.

[5] Presently, missing schedules and incomplete data are identified only at the time of the final consistency testing which occurs after data abstracting is complete. This contributes to processing delays.

[6] Report on Statistical Uses of Administrative Records, Statistical Policy Working Paper 6, Subcommittee on Statistical Uses of Administrative Records, Federal Committee on Statistical Methodology, Office of Federal Statistical Policy and Standards, U.S. Department of Commerce, December 1980.

[7] Longitudinal designs which include the same sample returns in the sample each year are utilized to maintain the reliability of estimates of year to year changes.

[8] While the resultant increase in the stability of the industry estimates would facilitate certain kinds of year-to-year comparisons, it could also mask the effect of bonafide changes in industrial activity in a given year. This would also occur if industry codes for given businesses were reassessed only periodically e.g., once every five years.

[9] The IRS Taxpayer Compliance Measurement Program (TCMP) compiles statistics on the results of comprehensive audits of taxpayers based on representative samples of various classes or types of income tax returns in order to estimate the total potential effects of audit. TCMP results might thus be used to "update" SOI, which is based on unaudited data.

[10] Left unresolved for purposes of this paper is the means by which the Statistics Division will be able to provide corporation data on an expedite basis to the Department of Commerce for use in benchmarking the national accounts in July of each year. Formerly, this need has been met by emphasizing the same early cutoff of the SOI sample used for the preliminary SOI statistics. With elimination of the preliminary statistics, the timing of the cutoff may be revised to a later date for the SOI complete statistics. This may prove incompatible with Commerce needs.

FIGURE 1.--INTERNAL REVENUE SERVICES - GEOGRAPHIC LOCATIONS

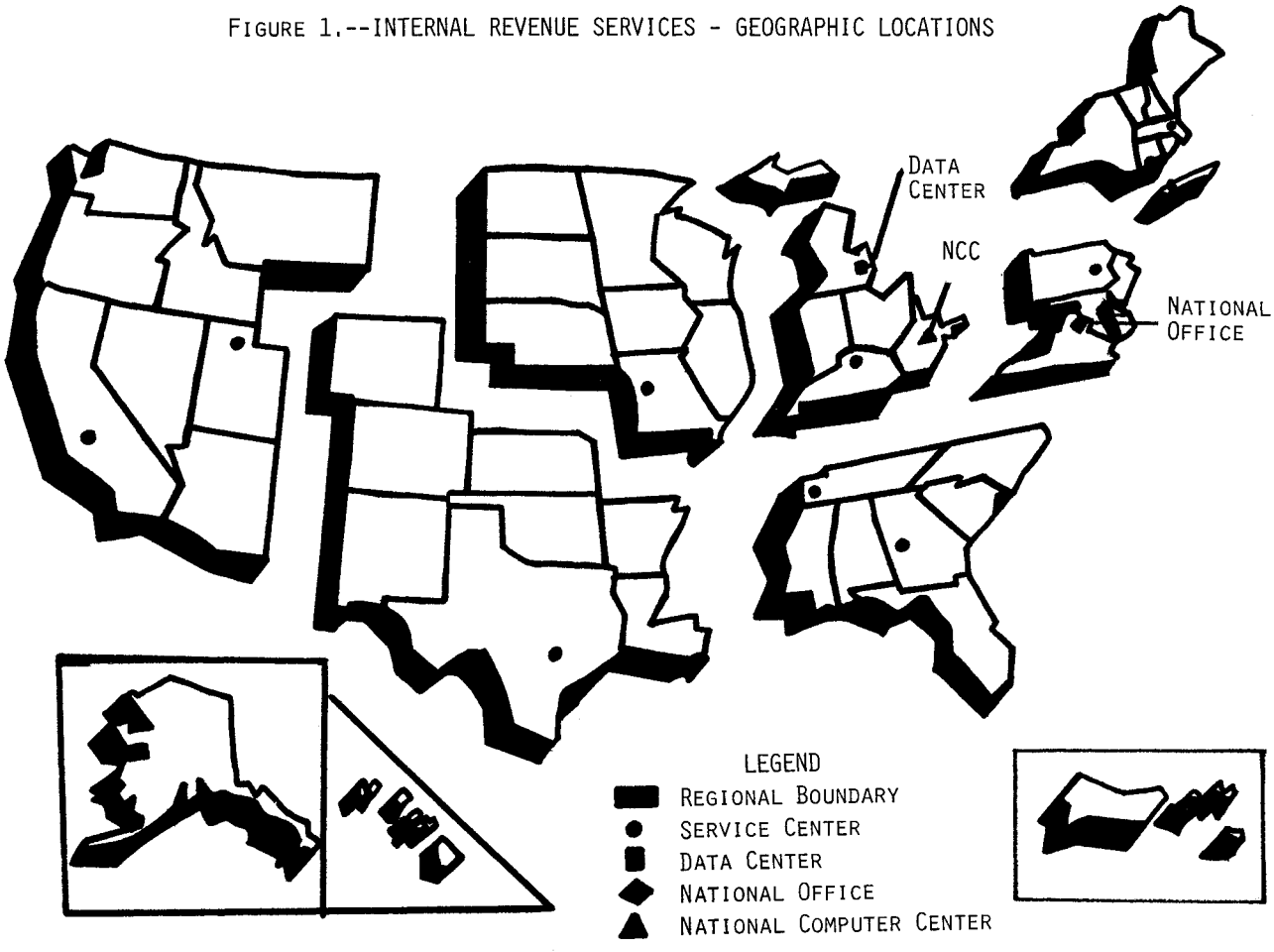

Figure 2.--Number of Returns Included in Statistics of Income Samples, by Tax Year

Program	Tax year						
	1979	1980	1981	1982	1983	1984	1985
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Individuals, total[1]	204.0	168.0	132.4	127.4	122.4	117.4	112.4
Nonbusiness	121.2	96.0	76.8	73.9	71.0	68.1	65.2
Business	82.8	72.0	55.6	53.5	51.4	49.3	47.2
Partnerships	50.0	40.0	35.0	35.0	35.0	35.0	35.0
Corporations:							
Sample, transaction tape	108.0	104.0	200.0	200.0	200.0	200.0	200.0
Subsample, total	77.6	90.0	95.0	95.0	95.0	95.0	95.0

[1]The size of the statistical sample for tax years beyond 1981 may be increased if unit processing costs can be reduced through revised methods.